U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

2.º
CICLO

FCUP
2017

U.PORTO

Regressão logística em dados com eventos raros

Ricardo Jorge Martins dos Santos

FC

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

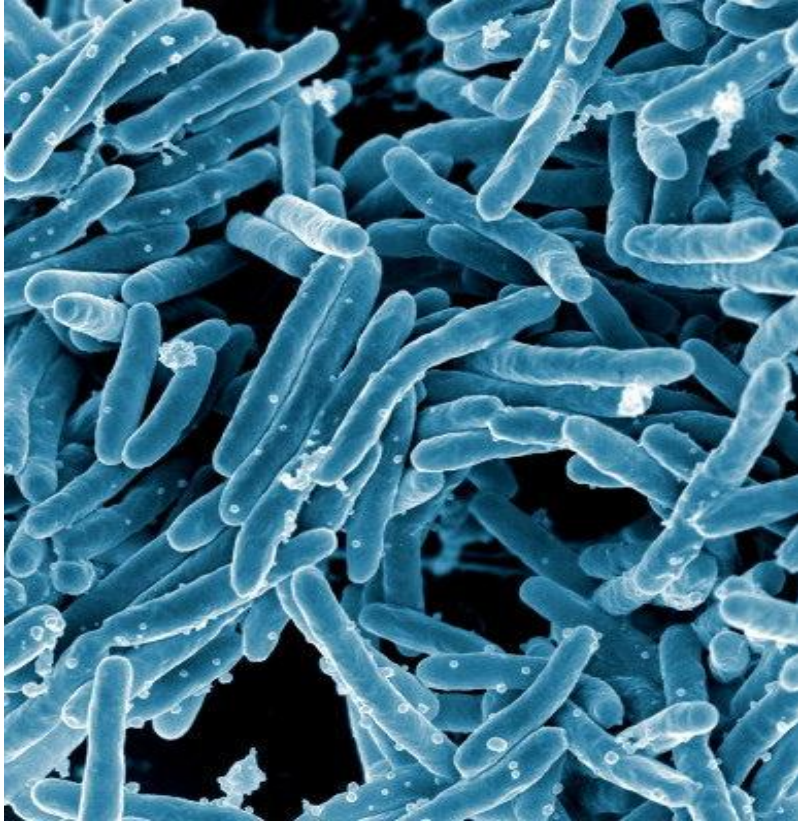
Regressão logística em dados com eventos raros

Ricardo Jorge Martins dos Santos

Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do Porto em
Engenharia Matemática

2017

MSc



Regressão logística em dados com eventos raros

Ricardo Jorge Martins dos Santos
Engenharia Matemática
Departamento de Matemática
2017

Orientador
Ana Rita Pires Gaio, Professora auxiliar, FCUP

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

... à memória da professora Rita.

Agradecimentos

As palavras que se seguem serão certamente poucas para expressar o meu agradecimento a quem contribuiu para a realização desta dissertação e do ciclo que com ela se encerra.

Quero começar por agradecer à Prof.ª Doutora Rita Gaio pela paciência, disponibilidade e apoio sem os quais esta dissertação não teria sido escrita. Pelos estímulos e ensinamentos que me transmitiu, o meu mais sincero obrigado!

Agradeço ao Dr. Bernardo Gomes e à Doutora Raquel Duarte pela disponibilização dos dados e pela sua disponibilidade e prontidão na resposta às dúvidas que foram surgindo ao longo da aplicação dos modelos, bem como dos resultados.

Agradeço aos meus pais, Adriano e Beatriz, por me terem permitido estudar na área que eu escolhi, sabendo eu que preferiam que tivesse seguido outra área. À minha "Mana"e ao meu cunhado Pedro agradeço-lhes as palavras de incentivo e a presença nos momentos mais importantes.

Aos meus tios Lurdes e Rui e à minha prima Mariana agradeço por me terem acolhido em sua casa durante estes últimos anos e por me terem aturado nos momentos de maior stress. Agradeço-lhes os cuidados e as preocupações, e peço-lhes desculpa pelos possíveis incómodos. Aos meus tios Ana e José Carlos e à minha prima Ângela agradeço os muitos jantares; os quais foram sempre momentos de descompressão.

Aos meus colegas e amigos, em especial à Raquel, à Maria e ao João, agradeço a presença, amizade, companheirismo e ajuda. Agradeço-lhes os momentos de descontração, sem os quais tudo teria sido bem mais difícil. Junto, neste agradecimento, as pessoas que conheci no Millennium bcp e aquelas com quem trabalho na Sonae, e que, de uma forma ou de outra, contribuíram para que eu crescesse enquanto profissional e ser humano.

A todos, muito obrigado!

Ricardo M. Santos

Resumo

Desde os trabalhos de Gauss (1777 - 1855) e Legendre (1752 - 1833) no século XIX, sobre o modelo linear clássico e o método dos mínimos quadrados, que os modelos de regressão se têm vindo a tornar cada vez mais uma ferramenta indispensável para estudar a relação entre uma variável resposta e uma ou mais variáveis explicativas.

De entre os inúmeros modelos de regressão existentes podemos destacar o modelo de regressão logística. Este modelo de regressão surgiu na primeira metade do século XX como resposta ao problema de modelar a relação entre uma variável resposta binária e uma ou mais variáveis explicativas. Na sequência das aplicações da metodologia levantou-se, aproximadamente meio século mais tarde, o problema da modelação de dados com eventos raros.

Para a estatística, um evento raro corresponde a uma variável aleatória binária para a qual o número de ocorrências do evento de interesse é muito inferior ao número de vezes em que este não ocorre. A raridade dos eventos constitui um problema ao nível da regressão logística. Na verdade, o desequilíbrio entre as duas categorias faz com que este modelo subestime a probabilidade de ocorrência do evento de interesse, sobrestimando portanto a sua não ocorrência.

Com o objetivo de tentar solucionar este problema estudam-se na presente dissertação, para além da regressão logística usual, a regressão logística condicional em estudos de caso-controlo, o modelo de regressão logística proposto por King e Zeng e a regressão logística de Firth. Os vários modelos são aplicados a dados reais sobre os desfechos do tratamento à Tuberculose em Portugal e um estudo comparativo é depois apresentado.

Conclui-se que o modelo que melhor se ajustou aos dados foi o da regressão logística condicional em estudos de caso-controlo.

Palavras-chave: Variável binária, regressão logística, eventos raros, estudos de caso-controlo, regressão logística condicional e tuberculose.

Abstract

Since the works of Gauss (1777 - 1855) and Legendre (1752 - 1833) in the XIX century, about the classical linear model and the least squares method, regression models have become more and more indispensable for the study of the association between a response variable and one or more explanatory variables.

Among the numerous existing regression models, we highlight the logistic regression model. This model appeared during the first half of the 20th century as a response to the problem of modeling the relationship between a binary response variable and one or more explanatory variables. Followed by applications of the methodology, problems of modeling data with rare events came to light approximately half a century later.

In Statistics, a rare event is associated with a binary random variable for which the number of occurrences of the event of interest is much lower than the number of times it does not occur. The rareness of these events is a problem in logistic regression. In fact, the imbalance between the frequencies of the two categories causes this model to underestimate the probability of occurrence of the event of interest, thus overestimating its non-occurrence.

With the purpose of trying to solve this problem, we study in the present thesis the conditional logistic regression model in case-control studies, the logistic regression model proposed by King and Zeng, and the Firth's logistic regression model. All models are applied to a real dataset concerning the outcome of the treatment to Tuberculosis in Portugal and a comparative study is presented. For completeness, the usual logistic regression is also described.

We conclude that the conditional logistic regression model in case-control studies is the one that best fits the data.

Keywords: Binary variable, logistic regression, rare events, case-control studies, conditional logistic regression and tuberculosis.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Desde os trabalhos de Gauss (1777 - 1855) e Legendre (1752 - 1833) no século XIX, sobre o modelo linear clássico e o método dos mínimos quadrados [McCullagh & Nelder, 1989], que os modelos de regressão se têm vindo a tornar cada vez mais uma ferramenta indispensável para estudar a relação entre uma variável resposta e uma ou mais variáveis explicativas. O modelo de regressão linear clássico é de longe o modelo de regressão mais utilizado [Hosmer & Lemeshow, 2000]. Neste modelo a variável resposta é contínua e depende linearmente de um conjunto de covariáveis.

Acontece que, tendo em conta as suas especificações, o modelo de regressão clássico não pode ser aplicado a todas as situações, desde logo pela natureza específica da resposta (figura 1.1) [Turkman & Silva, 2000].

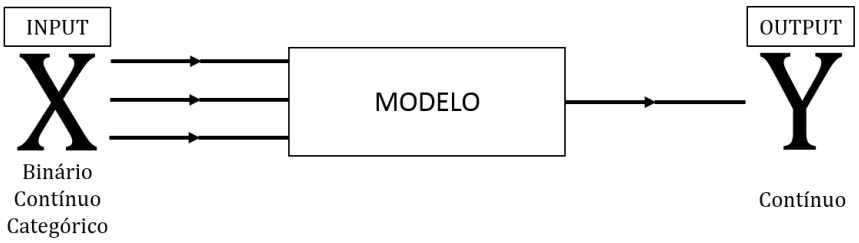


Figura 1.1: Esquema da regressão linear (baseado em [Wilson & Lorenz, 2015]).

Assim, ao longo dos últimos dois séculos foram sendo estudados novos modelos de regressão tendo como objetivo suprir as dificuldades do modelo de regressão linear clássico [McCullagh & Nelder, 1989].

Um desses modelos foi o modelo de regressão binomial (figura 1.2). Este modelo é utilizado para modelar a relação entre uma variável resposta binária e uma ou mais variáveis explicativas, que podem ser categóricas ou contínuas [Dobson, 2002]. É importante referir que existem vários modelos de regressão binomial, dependendo daquela que escolhemos como função de ligação. De entre estes modelos destacam-se a regressão binomial logística, a regressão binomial *probit* e a regressão binomial *log-log complementar*.

Na presente dissertação iremos abordar apenas o modelo de regressão binomial logística por este ser o mais aplicado. Referi-lo-emos apenas por modelo de regressão logística.

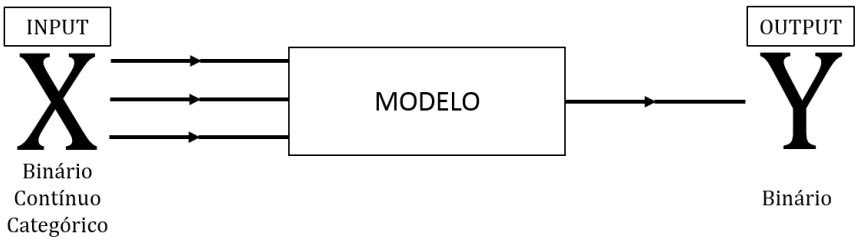


Figura 1.2: Esquema da regressão binomial (baseado em [Wilson & Lorenz, 2015]).

Entendemos aqui uma variável binária, ou dicotómica, como sendo uma variável aleatória que toma um de dois valores possíveis [Clayton & Hills, 2013]. Em geral, um destes valores representa o "sucesso"e o outro representa o "insucesso"de um certo acontecimento.

Acontece que dependendo do evento, este pode ocorrer com maior ou menor frequência. É neste ponto que se chega aquilo a que designamos eventos raros. Consideramos a ocorrência de um determinado evento como sendo um evento raro quando o número de vezes que este ocorre é consideravelmente menor do que o número de vezes em que este não ocorre.

Gary King e Langche Zeng [King & Zeng, 2001b] referem que há dois problemas na previsão de eventos raros utilizando as técnicas estatísticas usuais. Um desses problemas é que estas técnicas subestimam a probabilidade de acontecimento de um evento raro. A aplicação do modelo de regressão logística convencional a dados com eventos raros gera vieses nos coeficientes de regressão. Quanto maior o número de observações utilizadas para o ajustamento do modelo, maior é o vies gerado. Os autores acima indicados dizem-nos ainda que os vieses gerados são sempre na mesma direção, pelo que as probabilidades estimadas são sempre pequenas.

Com o objetivo de reduzir/corrigir estes vieses, foram aparecendo nas últimas décadas vários modelos de regressão para o estudo de casos associados a eventos raros. Ao longo da presente dissertação iremos abordar três destes modelos, sendo eles:

i. Regressão logística condicional em estudos de caso-controlo;

ii. Regressão logística com correção do viés, usando correção apriori e pesos; e

iii. Regressão logística de Firth.

Segundo N. E. Breslow [Breslow, 1996], os estudos de caso-controlo são usados para estudar doenças raras uma vez que estes são mais eficientes, quando comparados com estudos cohort. A regressão logística com correção do viés, usando correção apriori e pesos, foi proposta como forma de ultrapassar problemas com a baixa ocorrência do evento de interesse por King e Zeng em 2001, no artigo *Logistic Regression in Rare Events Data* [King & Zeng, 2001a]. Já a regressão logística de Firth apareceu no final do século XX e baseia-se na correção do viés, tendo em conta a função dos scores [Firth, 1993].

1.1 Objetivos da dissertação

Tendo em conta o tema abordado e a aplicação que se apresenta nesta dissertação, definimos os seguintes objetivos:

i. Estudar o modelo de regressão logística usual;

ii. Estudar a situação de eventos raros e abordar o fenómeno da separabilidade;

iii. Estudar modelos de regressão que visem responder aos problemas associados aos eventos raros; e

iv. Aplicar as metodologias estudadas a um conjunto de dados sobre a Tuberculose em Portugal.

1.2 Estrutura da dissertação

A presente dissertação encontra-se dividida em vários capítulos, sendo no início de cada capítulo feita uma breve menção aos temas que nele serão tratados.

No **capítulo 1** apresentamos uma introdução ao tema tratado na dissertação e formulamos os objetivos a que nos propusemos responder.

O modelo de regressão logística (usual), a função de verosimilhança e a estimação dos parâmetros de máxima verosimilhança são tratados no **capítulo 2**.

A definição de *eventos raros* é apresentada no **capítulo 3**. Neste capítulo para suporte da definição apresentamos dois exemplos de *eventos raros* em áreas distintas, e abordamos ainda o tema da separabilidade.

No **capítulo 4** são abordados os estudos de caso-controlo, os efeitos de confundimento e a regressão logística condicional em estudos de caso-controlo emparelhados.

Nos **capítulos 5** e **6** são apresentados, respectivamente, o modelo de regressão proposto por King e Zeng ([King & Zeng, 2001a]), no qual é feita a correção do viés nas situações em que a amostragem é feita usando emparelhamento, e a regressão de Firth, proposta por David Firth em 1993.

Apresentamos, no **capítulo 7**, uma aplicação dos vários modelos anteriormente indicados, utilizando dados reais sobre a Tuberculose em Portugal.

É aqui de relevo salientar que ao logo da dissertação, no final das secções nas quais são descritos modelos de regressão, são apresentadas as funções e respetivas bibliotecas que em R podem ser utilizadas no ajustamento de tais modelos.

No **capítulo 8** são apresentadas as principais conclusões, deixando-se algumas propostas para trabalhos futuros.

Capítulo 2

Modelo de Regressão logística

No presente capítulo falaremos do modelo de regressão logística, fazendo pequenos apontamentos às diferenças com o modelo de regressão linear. Abordaremos a função de verosimilhança e de como estimar os parâmetros do modelo de regressão que permitem maximizar a função de verosimilhança.

2.1 Introdução

O modelo de regressão logística apareceu em meados do século XX numa publicação de Dyke e Patterson. Nesta publicação os autores utilizaram a regressão logística para analisar um conjunto de dados sobre a proporção de indivíduos que tinham bons conhecimentos sobre cancro. O modelo de regressão logística tinha sido já utilizado por Berkson no contexto de um bioensaio no ano de 1944 [McCullagh & Nelder, 1989].

Tal como outros modelos estatísticos de regressão, o modelo de regressão logística apareceu para responder a problemas para os quais o modelo de regressão linear clássico não podia ser aplicado [Turkman & Silva, 2000], destacando-se, desde logo, os problemas com variável resposta binária [Hosmer & Lemeshow, 2000].

Segundo Harrell [Harrell, 2001] nas áreas da medicina e epidemiologia é comum o estudo de variáveis respostas binárias. Este dá-nos como exemplo o estudo da presença ou ausência de uma determinada doença. Em geral, nestes estudos, quer-se avaliar as relações entre as variáveis explicativas e a variável resposta.

2.2 Modelo de regressão logística

Comecemos por considerar o modelo mais simples, isto é, o modelo no qual a resposta binária está dependente apenas de uma variável explicativa.

Consideremos Y uma variável resposta binária, tal que:

$$Y = \begin{cases} 1, \text{ denota a ocorrência do evento de interesse;} \\ 0, \text{ denota a não ocorrência do evento de interesse.} \end{cases} \tag{2.1}$$

Nota: Na literatura é frequente designar a ocorrência de um evento de interesse por *sucesso* e a sua não ocorrência por *insucesso*.

Lembrando o modelo de regressão linear usual, quando a variável resposta é contínua podemos expressar a média condicional, $E(Y|x)$, por uma equação linear em x, tal que:

$$E(Y|x) = \beta_0 + \beta_1 x, \tag{2.2}$$

onde $\beta_0, \beta_1 \in \mathbb{R}$.

Note-se na expressão anterior que se x tomar um qualquer valor no intervalo entre $-\infty$ e $+\infty$, então também *o valor esperado de* $Y|x$ pode tomar um valor no intervalo $]-\infty, +\infty[$. Assim, é óbvio que, quando a variável resposta é binária a expressão (2.2) não pode ser aplicada [Cox, 1958]. Atente-se que nesta situação o valor da média condicional não pode ser maior que 1, nem menor que 0, isto é, quando Y é dicotómica, obrigatoriamente, tem de ser verificar que:

$$0 \leqslant E(Y|x) \leqslant 1.$$

Note-se ainda que, tal como referem King e Zeng [King & Zeng, 2001a], a média de uma variável binária corresponde à frequência relativa da ocorrência de eventos de interesse nos dados. Assim, podemos ver $E(Y|x)$ como a frequência relativa da ocorrência de eventos de interesse, nas observações que têm x como valor da variável explicativa.

De modo a simplificar a notação, consideremos $\pi(x) = E(Y|x)$ para representar *o valor esperado de* Y*, dado o valor de* x, no modelo de regressão logística.

Tendo em conta que o valor esperado de $Y|x$ está contido no intervalo $[0, 1]$, surge a necessidade de considerar uma transformação, $g(x)$, que relacione a média condicional, $\pi(x)$, com o preditor linear (o termo da direita na equação (2.2)), isto é:

$$g(x) = \beta_0 + \beta_1 x, \tag{2.3}$$

de modo a que o termo da esquerda possa também tomar valores em $\mathbb{R}$.

Aplicando a transformada logarítmica (de base e) a $\dfrac{\pi(x)}{1-\pi(x)}$, isto é, fazendo-se

$$g(x) = log\left(\frac{\pi(x)}{1-\pi(x)}\right) \tag{2.4}$$

conseguimos garantir que os termos (da direita e da esquerda) da equação (2.3) variam no mesmo intervalo, $\mathbb{R}$. Esta transformação é geralmente designada por **transformação logit** sendo representada por

$$g(x) = logit\left\{\pi(x)\right\} = log\left(\frac{\pi(x)}{1-\pi(x)}\right). \tag{2.5}$$

É aqui essencial fazer menção de que a função *logit* não é a única função de ligação que podemos usar como $g(x)$. Dada a sua relevância, são ainda de destacar pelo menos, as transformações *probit* e *log-log complementar*.

A transformação *probit* corresponde ao uso da função $probit\{\pi(x)\} = \Phi^{-1}\{\pi(x)\}$, onde a função $\Phi(\cdot)$ corresponde à função de distribuição de uma variável aleatória $N(0,1)$. Ao modelo de regressão binomial que se obtém utilizando como função de ligação a função *probit* designamos por modelo de regressão *probit*. Já a função *log-log complementar* corresponde ao uso de $g(x) = log(-log(1-\pi(x)))$. É de interesse referir que esta função corresponde à inversa da função de distribuição de Gumbel. A este modelo binomial costumamos designar por modelo de regressão log-log complementar .

$\star\star\star$

Odds e odds ratio:

Designamos por **odds** [1] de um evento o quociente entre a probabilidade da ocorrência do evento e a probabilidade da sua não ocorrência, isto é, é a razão entre duas probabilidades complementares [Stare & Maucort-Boulch, 2016]. Se representarmos a probabilidade do evento ocorrer por π, então o odds respetivo é dado pela seguinte expressão:

$$Odds = \frac{\pi}{1-\pi}.$$

Note-se, deste modo, que a transformada logit não é mais do que o logaritmo de um odds. É ainda fácil concluir da expressão (2.5) que o valor do odds é igual à exponencial do preditor linear, $g(x)$.

De um modo meramente ilustrativo, atente-se na tabela seguinte.

	$Evento$	$\overline{Evento}$
Grupo I	a	b
Grupo II	c	d

Tabela 2.1: Probabilidades de ocorrer, ou não, um determinado evento em diferentes Grupos, tal que $0 \leqslant a,\ b,\ c,\ d \leqslant 1$, $a+b=1$ e $c+d=1$.

Considerando π_1 e π_2, respetivamente, as probabilidades de para os grupos I e II ocorrer o evento de interesse, podemos calcular dois odds: o odds do evento no Grupo I ($Odds_1$) e o odds do evento no Grupo II ($Odds_2$).

$$Odds_1 = \frac{\pi_1}{1-\pi_1} = \frac{a}{b} \quad e \quad Odds_2 = \frac{\pi_2}{1-\pi_2} = \frac{c}{d}$$

Daqui segue que a frequência absoluta da ocorrência do evento de interesse é, no grupo I, $Odds_1 = a/b$ vezes a frequência absoluta da sua não ocorrência.

[1] A palavra odds vem da língua inglesa, não tendo uma tradução literal para o português; salienta-se ainda que esta palavra não possui plural.

Por seu turno, definimos como sendo um **odds ratio** (OR) o quociente entre dois odds, o que tendo em conta a tabela 2.1, resume-se à seguinte expressão:

$$OR = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc}$$

Deste modo conclui-se que o odds ratio do evento de interesse para os Grupos I e II é ad/bc. Este valor indica-nos que o evento de interesse ocorrerá ad/bc vezes no Grupo I, quando ocorre apenas uma vez para no Grupo II.

$\star\star\star$

Voltando à transformação logit, se considerarmos as expressões (2.3) e (2.4) temos:

$$log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x$$
$$\Leftrightarrow$$
$$\frac{\pi(x)}{1-\pi(x)} = exp(\beta_0 + \beta_1 x).$$

que, após multiplicarmos por $1-\pi(x)$ em ambos os termos, podemos simplificar para a expressão

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{2.6}$$

Tal como nos indica [Turkman & Silva, 2000], a função $F : \mathbb{R} \to [0,1]$, definida por

$$F(x) = \frac{e^x}{1+e^x} \tag{2.7}$$

é uma função de distribuição. Esta é, com efeito, a função de distribuição logística. É por esta razão que o modelo binomial com função de ligação *logit* é conhecido por modelo de regressão logística.

Uma diferença importante entre o modelo de regressão logística e o modelo de regressão linear corresponde à distribuição condicional da variável resposta. No modelo de regressão linear assume-se que a variável resposta pode ser expressa por $y = E(Y|x) + \epsilon$, onde a quantidade ϵ é designada por *erro* e expressa o desvio da observação em relação à média condicional. Geralmente assume-se que ϵ segue uma distribuição normal de média zero e variância constante. Isto leva a que a distribuição condicional que é assumida para a variável resposta, dado o valor de x, seja a normal com média $E(Y|x)$ e com variância constante. Note-se que tal não pode ser aplicado ao modelo de regressão logística.

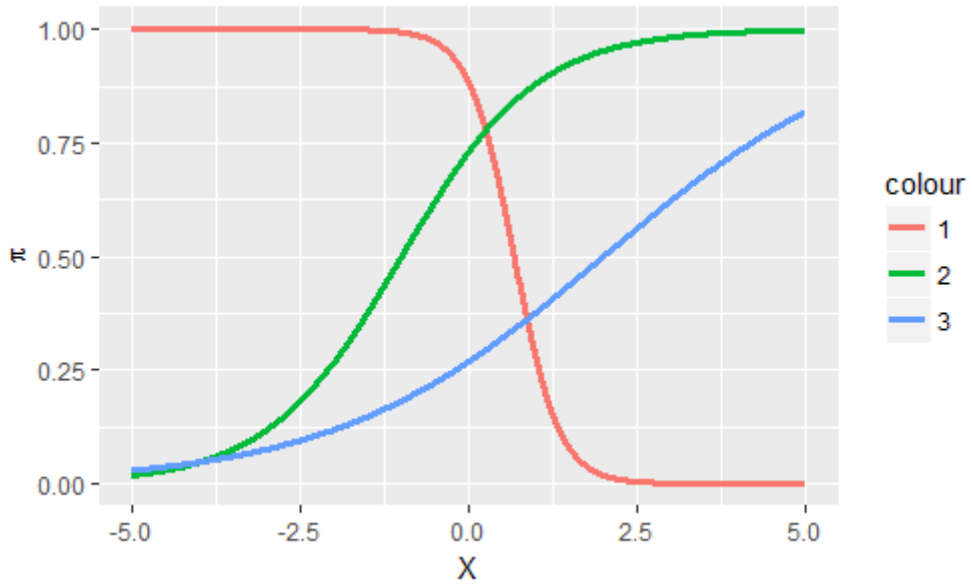


Figura 2.1: Representação gráfica de três curvas logísticas. A curva 1 foi obtida usando $\beta_0 = 2$ e $\beta_1 = -3$; a curva 2 utilizando $\beta_0 = 1$ e $\beta_1 = 1$; e a curva 3 utilizando $\beta_0 = -1$ e $\beta_1 = 0.5$.

Como a nossa variável resposta é binária, podemos expressar o seu valor, dado x, como sendo $y = E(Y|x) + \epsilon$, onde ϵ pode assumir um de dois valores possíveis. Se $y = 1$ então $\epsilon = 1 - \pi(x)$ com probabilidade $\pi(x)$, e se $y = 0$ então $\epsilon = -\pi(x)$ com probabilidade $1 - \pi(x)$. Assim, ϵ segue uma distribuição de média zero e variância $\pi(x)(1 - \pi(x))$. Isto é, tal como nos dizem Hosmer e Lemeshow, temos que a distribuição condicional da variável resposta segue uma distribuição binomial de média $\pi(x)$.

Depois de vermos o modelo de regressão simples, vamos agora abordar o modelo múltiplo, que não é mais do que uma generalização do primeiro.

Comecemos por considerar uma coleção de p valores referentes a p variáveis explicativas, denotados por $x^T = (x_1, x_1, ..., x_p)$. Considere-se ainda que a probabilidade condicional do sucesso ocorrer, dado um certo vetor x, para a variável resposta é $P(Y = 1|x) = \pi(x)$ e a probabilidade de ocorrer o insucesso é $P(Y = 0|x) = 1 - \pi(x)$.

O logit do modelo de regressão logística multipla é dado pela expressão

$$logit\{\pi(x)\} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p,$$

onde $\beta_0, \beta_1, ..., \beta_p \in \mathbb{R}$, o que é equivalente a

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} = \frac{e^{\beta_0 + \beta_1 x_1 + ... + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + ... + \beta_p x_p}}. \tag{2.8}$$

Tal como no modelo de regressão logística simples, as covariáveis $X_1, X_2, ..., X_p$ podem ser categóricas ou contínuas. As variáveis categóricas terão de ser incluídas no preditor linear através das suas variáveis indicatrizes [2]. Estas são variáveis acessórias que são criadas para

[2]Em inglês designadas *dummy variables*.

representar as várias categorias de uma certa variável categórica. De um modo geral, se a variável categórica possui k valores possíveis, isto é, k categorias, então são consideradas $k-1$ variáveis indicatrizes. Tal deve-se ao facto de o modelo apresentar um termo constante, no qual já será refletida uma das k categorias.

Tanto no modelo de regressão logística simples como no modelo de regressão logística múltiplo socorremo-nos da função do verosimilhança para determinar os parâmetros β's que melhor ajustam o modelo.

Ao nível do R: Para a obtenção do modelo de regressão logística usual no R podemos utilizar a função **glm( )** da biblioteca **stats**. Esta função é usada para a obtenção dos vários modelos lineares generalizados, pelo que é essencial definirmos "*family=binomial(link="logit")*"para obtermos o modelos de regressão logística. A função apresenta ainda um variado conjunto de parâmetros que podem ser alterados, de modo a obtermos o que pretendemos. Apresenta-se de seguida a função e os respetivos parâmetros:

```
1 glm(formula, family=binomial(link="logit"), data, weights, subset,
2     na.action, start=NULL, etastart, mustart, offset, control=list(...),
3     model=TRUE, method="glm.fit", x=FALSE, y=TRUE, contrasts=NULL, ...)
```

Mais informações sobre esta função podem ser encontradas no manual do R, em `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/glm.html`.

2.3 Função de verosimilhança

Comecemos por considerar $Y^T = (Y_1, Y_2, ..., Y_n)$ um vetor aleatório e $f(y;\theta)$ a função de probabilidade conjunta de Y, com $\theta^T = (\theta_1, \theta_2, ..., \theta_p)$.

Algebricamente a função de verosimilhança, $L(\theta; y)$, e a função densidade de probabilidade (f.d.p.) $f(y;\theta)$ são iguais. A troca de notação justifica-se com o facto de na f.d.p. os valores do vetor θ serem fixos, variando os valores do vetor y, e na função de verosimilhança, os valores do vetor y serem fixos, variando os valores do vetor θ [Dobson, 2002].

$$\begin{aligned} L(\theta; y) &= L(\theta; Y_1, Y_2, ..., Y_n) \qquad (= f(y;\theta)) \\ &= f(Y_1, Y_2, ..., Y_n; \theta) \\ &= f(Y_1;\theta) f(Y_2;\theta) ... f(Y_n;\theta) \qquad\qquad (Y_i's \ i.i.d.'s) \\ &= \prod_{i=1}^{n} f(Y_i;\theta) \end{aligned} \tag{2.9}$$

Considerando Ω o espaço paramétrico formado por todos os valores que pode tomar o vetor de parâmetros θ e $\widehat{\theta}$ o estimador de máxima verosimilhança do vetor θ, então tem de se verificar:

$$L(\widehat{\theta}; y) \geq L(\theta; y), \ para \ todo \ \theta \in \Omega. \tag{2.10}$$

Note-se que se $\widehat{\theta}$ é o vetor que maximiza a função de verosimilhança este é ao mesmo tempo o vetor que maximiza a função de log-verosimilhança,

$$l(\theta; y) = log(L(\theta; y))$$

uma vez que a função logarítmica é monótona. Assim, à semelhança de (2.10), temos que

$$l(\widehat{\theta}; y) \geq l(\theta; y), \ para \ todo \ \theta \in \Omega.$$

De um modo geral considera-se a função de log-verosimilhança uma vez que esta é mais fácil de trabalhar que a função de verosimilhança.

Teoricamente obtemos o estimador de máxima verosimilhança $\widehat{\theta}$ derivando a função de log-verosimilhança em ordem a cada um dos elementos θ_j do vetor θ e resolvendo de seguida as equações de verosimilhança,

$$\frac{\partial l(\theta; y)}{\partial \theta_j} = 0 \quad for \ j = 1, ..., p.$$

Após ser determinada uma potencial solução $\widehat{\theta}$ é necessário verificar que a matriz das segundas derivadas,

$$\frac{\partial^2 l(\theta; y)}{\partial \theta_j \partial \theta_k},$$

é definida negativa para tal solução, de modo a garantirmos que tal corresponde a um máximo.

Para além de verificar que, para $\theta = \widehat{\theta}$, a matriz das segundas derivadas é definida negativa, é ainda necessário verificar que $\widehat{\theta}$ é mesmo o máximo absoluto. Assim, é necessário determinar

todos os candidatos a máximos absolutos, isto é, todos os θ's para os quais a matriz hessiana é definida negativa e verificar qual deles é o máximo absoluto. O parâmetro θ que assim determinarmos é aquele que maximiza a verosimilhança, pelo que corresponde a $\widehat{\theta}$.

Ao aplicarmos a função de verosimilhança para a estimação do vetor de parâmetros $\widehat{\beta}^T = (\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_p)$ estamos a determinar o vetor de parâmetros que maximiza a probabilidade de se obter o conjunto de dados observados.

Podemos, tendo em conta a f.d.p. da binomial, resumir o contributo de um qualquer par da forma (x_i, y_i) para a função de verosimilhança utilizando a seguinte expressão:

$$\pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}.$$

Considerando que dispomos de n observações independentes para a estimação dos parâmetros do modelo, o que corresponde a n pares da forma (x_i, y_i), então a função de verosimilhança é dada pela seguinte expressão:

$$L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}, \tag{2.11}$$

onde β é o vetor dos parâmetros a determinar.

Aplicando a função logaritmo à função verosimilhança da expressão (2.11):

$$l(\beta) = log(L(\beta)) = \sum_{i=1}^{n} [y_i log(\pi(x_i)) + (1 - y_i) log(1 - \pi(x_i))]. \tag{2.12}$$

Como o objetivo é determinar o vetor β para o qual a função de verosimilhança é máxima, o primeiro passo corresponde a derivar a função de log-verosimilhança em ordem a cada β_j, $j = 0, 1, ..., p$, do vetor β, e igualar a derivada a zero. Como há $p + 1$ parâmetros então obteremos $p + 1$ equações de verosimilhança, que correspondem ao seguinte sistema:

$$\begin{cases} \sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \\ \sum_{i=1}^{n} [y_{ij} - \pi(x_{ij})] = 0, \quad j = 1, 2, ..., p \end{cases} \tag{2.13}$$

É de se referir que os valores dos parâmetros β's que são solução do sistema (2.13) são chamados de **estimadores de máxima verosimilhança** e são denotados por $\widehat{\beta}^T = (\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_p)$.

Chegados a este ponto, é de se notar que as equações do sistema em (2.13) não são lineares. Deste modo para podermos chegar a uma solução teremos de nos socorrer de métodos especiais para a resolução do sistema [Hosmer & Lemeshow, 2000]. Tais métodos são métodos numéricos iterativos.

Um método iterativo para a obtenção da solução do sistema (2.13) é o método **iterativo dos mínimos quadrados ponderados**, também conhecido por I.W.L.S. (sigla de iterative weighted least squares), que será apresentado na secção seguinte.

2.4 Métodos iterativos para a estimação dos parâmetros que maximizam a função de verosimilhança

Um método iterativo é um procedimento que visa estimar a solução de um determinado problema, socorrendo-se de iterações sucessivas que geram uma sequência de aproximações de tal solução. Um método iterativo bastante conhecido e amplamente utilizado para a estimação de raízes é o **Método de Newton-Raphson**.

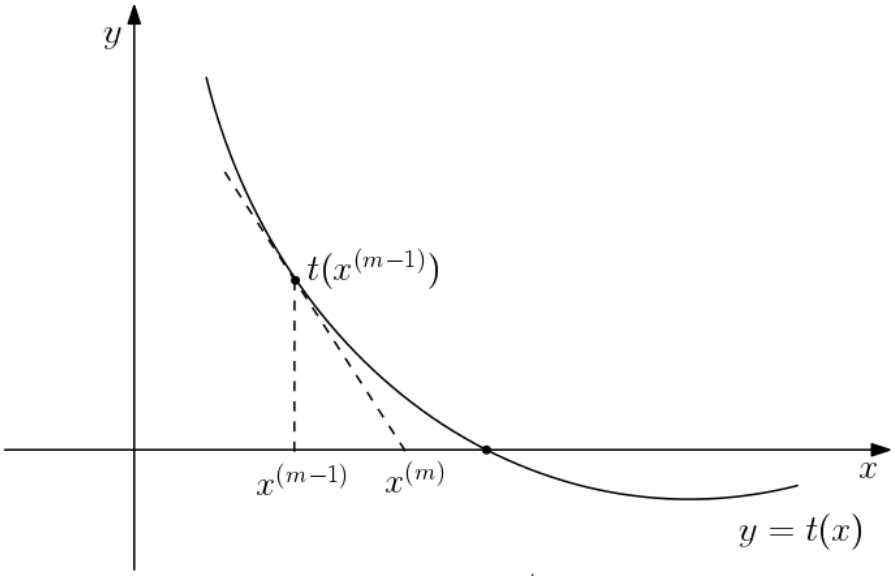


Figura 2.2: Método de Newton-Raphson para determinar a solução da equação $t(x) = 0$.

No método de Newton-Raphson queremos determinar o valor de x para o qual a função $t(x)$ cruza o eixo das abcissas, isto é, queremos solucionar $t(x) = 0$.

O declive da função $t(x)$ no ponto de abcissa $x^{(m-1)}$ é dado pela expressão

$$\left[\frac{dt(x)}{dx}\right]_{x=x^{(m-1)}} = t'(x^{m-1}) = \frac{t(x^{(m-1)}) - t(x^{(m)})}{x^{(m-1)} - x^{(m)}}$$

onde a distância $x^{(m-1)} - x^{(m)}$ é pequena.

Assim, se $x^{(m)}$ é a solução que pretendemos, ou seja, se $x^{(m)}$ é uma aproximação boa o suficiente da solução que procurávamos, então podemos reescreve-la da seguinte forma:

$$x^{(m)} = x^{(m-1)} - \frac{t(x^{(m-1)})}{t'(x^{(m-1)})}. \tag{2.14}$$

Chegados a este ponto é fundamental ter em atenção que para fazer o método convergir necessitamos de lhe fornecer uma aproximação inicial da solução, a qual representamos por $x^{(0)}$.

Consideremos agora uma variável aleatória Y cuja log-verosimilhança é dada pela expressão:

$$l(\theta; y) = log(L(\theta; y)),$$

e consideremos que esta segue uma distribuição pertencente à família exponencial, isto é, a sua função de densidade de probabilidade pode ser escrita da seguinte forma [Dobson, 2002]:

$$f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)} \tag{2.15}$$

onde a, b, s e t são funções conhecidas. Note-se que esta expressão é equivalente a:

$$f(y;\theta) = e^{a(y)b(\theta)+c(\theta)+d(y)}, \tag{2.16}$$

se considerarmos $s(y) = e^{d(y)}$ e $t(\theta) = e^{c(\theta)}$.

Designamos por U a **função dos scores** que corresponde à função resultante da derivação da função de verosimilhança em ordem a θ, isto é:

$$U(\theta;y) = \frac{dl(\theta;y)}{d\theta}. \tag{2.17}$$

Note-se que, determinar os parâmetros que maximizam a função de log-verosimilhança, $l(\theta;y)$, isto é, os estimadores de máxima verosimilhança $\widehat{\theta}$, é o mesmo que determinar os zeros da função U. Queremos com isto fazer notar que os estimadores de máxima verosimilhança $\widehat{\theta}$ são a solução da equação $U(\theta;y) = 0$.

Assim, na expressão do método de Newton-Raphson devemos substituir x por θ e a função $t(x)$ pela $U(\theta)$, ou seja, ficamos com a seguinte expressão:

$$\theta^{(m)} = \theta^{(m-1)} - \frac{U(\theta^{(m-1)})}{U'(\theta^{(m-1)})}. \tag{2.18}$$

Para a estimação de máxima verosimilhança é usual considerar uma aproximação de U' pelo seu valor esperado, $E[U']$.

Deste modo, vamos de seguida determinar tal valor esperado, com o objetivo de reescrever a expressão anterior utilizando a aproximação.

Assim, note-se que seguindo Y uma distribuição pertencente à família exponencial, então podemos escrever a sua função de log-verosimilhança da seguinte forma:

$$l(\theta;y) = a(y)b(\theta) + c(\theta) + d(y). \tag{2.19}$$

Derivando a expressão anterior em ordem a θ, temos:

$$U(\theta;y) = \frac{dl(\theta;y)}{d\theta} = a(y)b'(\theta) + c'(\theta)$$

Como a **estatística score**, $U(\theta;y)$, depende de y, podemos escreve-la como uma variável aleatória, tal que:

$$U = a(Y)b'(\theta) + c'(\theta) \tag{2.20}$$

então o seu valor esperado pode ser escrito da forma:

$$E[U] = b'(\theta)E[a(Y)] + c'(\theta). \tag{2.21}$$

Note-se que sendo $f(y;\theta)$ a f.d.p. da nossa variável Y então, por definição,

$$\int f(y;\theta)dy = 1,$$

pelo que, ao derivarmos ambos os termos em ordem a θ vamos obter

$$\frac{d}{d\theta}\int f(y;\theta)dy = \frac{d}{d\theta}.1 = 0 \tag{2.22}$$

Da expressão anterior segue que, sendo a ordem de integração e de derivação no termo da esquerda invertida, então podemos escrever a expressão (2.22) da forma

$$\int \frac{df(y;\theta)}{d\theta}dy = 0.$$

Seguindo o mesmo pensamento, podemos escrever uma segunda derivada da seguinte forma:

$$\int \frac{d^2f(y;\theta)}{d\theta^2}dy = 0. \tag{2.23}$$

Assim, como Y segue uma distribuição pertencente à família exponencial temos

$$\frac{df(y;\theta)}{d\theta} = [a(y)b'(\theta) + c'(\theta)]f(y;\theta),$$

pelo que,

$$\int [a(y)b'(\theta) + c'(\theta)]f(y;\theta)dy = 0,$$

o que, tendo em conta a definição de valor esperado, nos permite concluir o valor esperado de $a(Y)$:

$$b'(\theta)E[a(Y)] + c'(\theta) = 0 \quad \Leftrightarrow \quad E[a(Y)] = -\frac{c'(\theta)}{b'(\theta)}. \tag{2.24}$$

Ao substituir a expressão (2.24) na expressão (2.21) podemos concluir o valor esperado de U

$$E[U] = b'(\theta)\left[-\frac{c'(\theta)}{b'(\theta)}\right] + c'(\theta) = 0. \tag{2.25}$$

De um modo idêntico podemos concluir a variância de U, $var[U]$. Note-se que

$$var[U] = var[a(Y)b'(\theta) + c'(\theta)] = [b'(\theta)]^2 var[a(Y)]. \tag{2.26}$$

Como

$$\frac{d^2f(y;\theta)}{d\theta^2} = [a(y)b''(\theta) + c''(\theta)]f(y;\theta) + [a(y)b'(\theta) + c'(\theta)]^2 f(y;\theta),$$

e, em especial podemos escrever,

$$[a(y)b'(\theta) + c'(\theta)]^2 f(y;\theta) = [b'(\theta)]^2\{a(y) - E(a(Y))\}^2 f(y;\theta)$$

então pela expressão (2.23) e por $var[a(Y)] = \int\{a(y) - E(a(Y))\}^2 f(y;\theta)$ temos que

$$\int \frac{d^2f(y;\theta)}{d\theta^2}dy = b''(\theta)E[a(Y)] + c''(\theta) + [b'(\theta)]^2 var[a(Y)] = 0. \tag{2.27}$$

Pelo que concluimos que a $var[a(Y)]$ é dada pela seguinte expressão:

$$var[a(Y)] = \frac{-b''(\theta)E[a(Y)] - c''(\theta)}{[b'(\theta)]^2} = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}. \tag{2.28}$$

Assim concluímos que a $var[U]$ é da forma:

$$var[U] = [b'(\theta)]^2 var[a(Y)] = \frac{b''(\theta)c'(\theta)}{b'(\theta)} - c''(\theta). \tag{2.29}$$

Posto isto, como $U' = \dfrac{dU}{d\theta} = a(Y)b''(\theta) + c''(\theta)$, temos que o seu valor esperado é dado pela seguinte expressão

$$\begin{aligned} E[U'] &= b''(\theta)E[a(Y)] + c''(\theta) \\ &= b''(\theta)\left[-\frac{c'(\theta)}{b'(\theta)}\right] + c''(\theta) \\ &= -var[U]. \end{aligned} \tag{2.30}$$

À variância de U, $var[U]$, damos o nome de **informação** e denotamos por $\mathfrak{I}$ [Dobson, 2002]. Logo $E[U'] = -\mathfrak{I}$.

Assim, tomando a expressão de U' pelo seu valor esperado, podemos escrever a equação de estimação do método de Newton-Raphson da seguinte forma:

$$\theta^{(m)} = \theta^{(m-1)} + \frac{U(\theta^{(m-1)})}{\mathfrak{I}(\theta^{(m-1)})}. \tag{2.31}$$

De um modo geral a esta variante do método de Newton-Raphson designamos por **método de Scoring**.

Se Y for uma variável aleatória que segue uma distribuição binomial com parâmetros n e π, $Y \sim B(n, \pi)$, a expressão da sua função densidade de probabilidade é dada pela expressão:

$$f(y;\pi) = \begin{pmatrix} \mathrm{n} \\ \mathrm{y} \end{pmatrix} \pi^y (1-\pi)^{n-y},$$

onde y toma os valores $0, 1, ..., n$.

Note-se que a expressão anterior é equivalente à que se segue:

$$\begin{aligned} f(y;\pi) &= exp\left\{ylog(\pi) - ylog(1-\pi) + nlog(1-\pi) + log\begin{pmatrix} \mathrm{n} \\ \mathrm{y} \end{pmatrix}\right\} \\ &= exp\left\{ylog\left(\frac{\pi}{1-\pi}\right) + nlog(1-\pi) + log\begin{pmatrix} \mathrm{n} \\ \mathrm{y} \end{pmatrix}\right\}, \end{aligned} \tag{2.32}$$

pelo que, considerando a expressão (2.16) e $\pi = \theta$ temos que: $a(y) = y$, $b(\theta) = b(\pi) = log\left(\dfrac{\pi}{1-\pi}\right)$, $c(\theta) = c(\pi) = nlog(1-\pi)$ e $d(y) = log\begin{pmatrix} \mathrm{n} \\ \mathrm{y} \end{pmatrix}$, isto é, Y segue uma distribuição

pertencente à família exponencial.

Como na regressão logística a variável dependente segue uma distribuição binomial, ao considerarmos um conjunto $Y_1, Y_2, ..., Y_n$ de variáveis aleatórias independentes, para cada Y_i podemos considerar a seguinte expressão da função de verosimilhança que lhe está associada:

$$l_i = y_i b(\theta_i) + c(\theta_i) + d(y_i). \tag{2.33}$$

Deste modo, a função de log-verosimilhança de todos os Y_i's é dada pela expressão:

$$l = \sum_{i=1}^{n} l_i = \sum y_i b(\theta_i) + \sum c(\theta_i) + \sum d(y_i), \tag{2.34}$$

onde as funções b,c e d são definidas como em (2.32), e tais que:

$$\begin{aligned} E[Y_i] &= \mu_i = -\frac{c'(\theta_i)}{b'(\theta_i)} \\ var[Y_i] &= \frac{b''(\theta_i)c'(\theta_i) - c''(\theta_i)b'(\theta_i)}{[b'(\theta_i)]^3} \\ e\ \eta_i &= x_i^T \beta, \end{aligned} \tag{2.35}$$

sendo x_i um vetor com os elementos $x_{ij},\ j = 1, ..., p$.

Assim, para obtermos os estimadores de máxima verosimilhança dos parâmetros β_j temos de derivar l em ordem a cada um dos β_j's. Tal derivação, aplicando-se a regra da cadeia, resulta em:

$$\frac{\partial l}{\partial \beta_j} = U_j = \sum_{i=1}^{n} \left[\frac{\partial l_i}{\partial \beta_j}\right] = \sum_{i=1}^{n} \left[\frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j}\right] \tag{2.36}$$

Tratando separadamente o termo mais à direita, temos:

$$\begin{aligned} \frac{\partial l_i}{\partial \theta_i} &= y_i b'(\theta_i) + c'(\theta_i) = b'(\theta_i)(y_i - \mu_i) \\ \frac{\partial \theta_i}{\partial \mu_i} &= \left[\frac{\partial \mu_i}{\partial \theta_i}\right]^{-1} = \left[\frac{b''(\theta_i)c'(\theta_i) - c''(\theta_i)b'(\theta_i)}{[b'(\theta_i)]^2}\right]^{-1} \\ &= \frac{1}{b'(\theta_i)var[Y_i]} \\ e\ \frac{\partial \mu_i}{\partial \beta_i} &= \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_i} = \frac{\partial \mu_i}{\partial \eta_i}x_{ij} \end{aligned} \tag{2.37}$$

Deste modo, juntando (2.36) e (2.37), temos que:

$$U_j = \sum_{i=1}^{n} \left[\frac{(y_i - \mu_i)}{var[Y_i]} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right]. \tag{2.38}$$

Daqui segue que a matriz de covariância, dada pela expressão $\mathfrak{I}_{jk} = E[U_j U_k]$, e tendo em conta a independência das variáveis Y_i's, tem a seguinte expressão:

$$\mathfrak{I}_{jk} = \sum_{i=1}^{n} \frac{E[(Y_i - \mu_i)^2]x_{ij}x_{ik}}{[var(Y_i)]^2} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2.$$

Como $E[(Y_i - \mu_i)^2] = var[Y_i]$, podemos simplificar a expressão anterior para

$$\mathfrak{I}_{jk} = \sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2. \tag{2.39}$$

À matriz $\mathfrak{I}$ damos o nome de **matriz de informação** [Dobson, 2002].

Chegados a este ponto podemos generalizar a expressão da estimação pelo método de scoring para

$$\beta^{(m)} = \beta^{(m-1)} + [\mathfrak{I}^{(\mathfrak{m}-\mathfrak{1})}]^{-1} U^{(m-1)},$$

onde $\beta^{(m)}$ é o vetor da estimação dos parâmetros $\beta_1, \beta_2, ..., \beta_p$ na m-ésima iteração.

Tendo em conta que $[\mathfrak{I}^{(\mathfrak{m}-\mathfrak{1})}]^{-1}$ representa a inversa da matriz de informação podemos reescrever a expressão anterior multiplicando pela matriz de informação de ambos os lados:

$$\mathfrak{I}^{(\mathfrak{m}-\mathfrak{1})}\beta^{(m)} = \mathfrak{I}^{(\mathfrak{m}-\mathfrak{1})}\beta^{(m-1)} + U^{(m-1)}. \tag{2.40}$$

Note-se agora que podemos escrever $\mathfrak{I}$ a partir da expressão (2.39) da seguinte forma:

$$\mathfrak{I} = X^T W X, \tag{2.41}$$

onde W é uma matriz diagonal de dimensão $n \times n$ com elementos $\omega_{ii} = \dfrac{1}{var(Y_i)} \left(\dfrac{\partial \mu_i}{\partial \eta_i}\right)$.

O termo da direita na expressão (2.40) resulta num vetor com elementos

$$\sum_{k=1}^{p}\sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \beta_k^{(m-1)} + \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{var[Y_i]} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i}\right),$$

a qual podemos escrever segundo a notação vetorial como

$$X^T W z, \tag{2.42}$$

se considerarmos z um vetor com elementos

$$z_i = \sum_{k=1}^{p} x_{ik}\beta_k^{(m-1)} + (y_i - \mu_i) \left(\frac{\partial \eta_i}{\partial \mu_i}\right).$$

Assim, a equação iterativa (2.40) pode ser escrita em notação vetorial da seguinte forma:

$$X^T W X \beta^{(m)} = X^T W z. \tag{2.43}$$

A esta generalização é dado o nome de método iterativo dos mínimos quadrados ponderados [Dobson, 2002].

2.5 Teste à significância dos parâmetros

Após obtermos uma estimação dos coeficientes do modelo de regressão a nossa preocupação passa a ser a avaliação da significância das variáveis no modelo. De um modo geral esta avaliação requer a realização de um teste de hipóteses [Hosmer & Lemeshow, 2000] que visa verificar se as variáveis explicativas no modelo estão "significativamente"relacionadas com a variável resposta.

Considerando uma hipotética variável explicativa X, uma abordagem comummente utilizada para testar a significância do coeficiente do modelo de regressão associado a esta variável passa pelo seguinte princípio:

> *Comparar os valores obtidos para a resposta usando os modelos de regressão com e sem a variável X.*

Na regressão logística a comparação entre os valores observados e previstos para a variável resposta é baseada na função de log-verosimilhança.

Para entender esta comparação é essencial conhecer o conceito de **modelo saturado**. Definimos um modelo saturado como sendo um modelo que tem tantos parâmetros quantas observações têm os dados. Este modelo carateriza-se por não apresentar erro aleatório, não ser informativo e por explicar toda a variabilidade dos dados.

A comparação dos valores observados com os valores previstos usando a função de log-verosimilhança é feita considerando-se a seguinte expressão:

$$D = -2log\left[\frac{\text{verosimilhança do modelo ajustado}}{\text{verosimilhança do modelo saturado}}\right]. \tag{2.44}$$

Na expressão anterior, o rácio entre as funções de verosimilhança é designado por razão de verosimilhanças. A aplicação do -2 e do log é importante para que a quantidade D siga uma distribuição conhecida, de modo a que lhe possamos aplicar testes de hipóteses.

De um outro modo, podemos escrever a expressão de D da seguinte forma:

$$D = -2\sum_{i=1}^{n}\left[y_i log\left(\frac{\widehat{\pi}_i}{y_i}\right) + (1-y_i)log\left(\frac{1-\widehat{\pi}_i}{1-y_i}\right)\right]. \tag{2.45}$$

Designaremos, à semelhança de McCullagh e Nelder (1983), a quantidade D por **desviância**.

No caso da regressão logística, como a variável resposta é binária, o valor da função de verosimilhança do modelo saturado é 1 [Hosmer & Lemeshow, 2000], isto é, pela definição de modelo saturado segue que $\widehat{\pi}_i = y_i$, logo:

$$l(\text{modelo saturado}) = \prod_{i=1}^{n} y_i^{y_i}(1-y_i)^{y_i} = 1.$$

Assim, no caso da regressão logística a expressão da desviância resume-se a:

$$D = -2log(\text{verosimilhança do modelo ajustado}). \tag{2.46}$$

Consideremos G a diferença entre a desviância do modelo sem a variável em estudo e a desviância do modelo com a variável, isto é,

$$G = D(\text{modelo ajustado sem a variável}) - D(\text{modelo ajustado com a variável}). \tag{2.47}$$

Note-se que o valor da log-verosimilhança do modelo saturado é comum aos dois valores de D. Deste modo, a equação (2.47) pode ser reescrita da seguinte forma:

$$G = -2log\left[\frac{(\text{modelo ajustado sem a variável})}{(\text{modelo ajustado com a variável})}\right]. \tag{2.48}$$

Baseando-nos na estatística G podemos fazer um teste de modo a avaliar se uma certa variável é ou não relevante para o modelo, isto é, se o modelo se encontra melhor ajustado com ou sem a variável em causa. Assim, podemos colocar as seguintes hipóteses:

$$\text{H0: } \beta_j = 0$$
$$\text{H1: } \beta_j \neq 0,\ j \in \{0, 1, ..., p\}$$

Tendo em conta que apenas estamos a testar a significância de uma certa variável no modelo, então a estatística G segue uma distribuição χ^2 com um grau de liberdade, isto é, $G \sim \chi^2(1)$.

Ao considerarmos um nível de significância α, rejeitamos a hipótese nula, H0, quando $G > \chi^2_{1-\alpha}(1)$ o que é equivalente a verificar $P(\chi^2(1) > G) < \alpha$. Se assim for aceitamos a hipótese alternativa, H1, como verdadeira para um nível de significância α.

Caso não se verifique que $G > \chi^2_{1-\alpha}(1)$, então não podemos rejeitar a hipótese nula, H0.

De uma forma mais genérica, podemos avaliar a significância de um conjunto de t variáveis explicativas, $\underline{\beta} = (\beta_{j_1}, \beta_{j_2}, ..., \beta_{j_t})^T, \{j_1, j_2, ..., j_t\} \subset \{1, 2, ..., p\}$. Neste caso, a expressão da estatística G corresponde a:

$$G = -2log\left[\frac{(\text{modelo ajustado sem as variáveis})}{(\text{modelo ajustado com as variáveis})}\right]. \tag{2.49}$$

Podemos colocar hipóteses da seguinte forma:

$$\text{H0: } \underline{\beta} = (0, 0, ..., 0)^T$$
$$\text{H1: } \underline{\beta} \neq (0, 0, ..., 0)^T$$

Tendo em conta que $\underline{\beta}$ tem t elementos, então rejeitaremos a hipótese nula, para um nível de significância α, se verificarmos $G > \chi^2_{1-\alpha}(t)$, aceitando a hipótese alternativa, H1, como verdadeira. Caso contrário não rejeitamos H0.

É de referir que t pode ser igual a p. Nesta situação estar-se-á a comparar o modelo que contém todas as variáveis, com o modelo que apenas possui o coeficiente constante.

Ao teste descrito dá-se o nome de **teste da razão de verosimilhanças**.

Com uma finalidade idêntica podemos aplicar o **teste de Wald** e o **teste de Score**.

O teste de Wald permite-nos avaliar se algum ou todos os coeficientes são não nulos, usando para tal uma estatística de teste, em geral denominada por **estatística de Wald**, que compara as estimativas de máxima verosimilhança dos parâmetros $\beta_i, i = 0, 1, ..., p$ com o seu erro padrão.

A expressão da estatística univariada é da forma:

$$W_j = \frac{\widehat{\beta}_j}{\widehat{SE}(\widehat{\beta}_j)} \tag{2.50}$$

onde $j = 0, 1, ..., p$ e SE diz respeito ao erro padrão.

A estatística de Wald univariada segue uma distribuição normal standard, $W_j \sim N(0, 1)$, se estivermos a considerar a nossa hipótese nula H0: $\beta_j = 0$ [Hosmer & Lemeshow, 2000]. Se $j = 0$ então estamos a calcular a estatística de teste para o coeficiente independente β_0 e estaremos a testar se o coeficiente é nulo ou não nulo. Por outro lado se $j = 1, 2, ..., p$ então estaremos a avaliar se alguma das variáveis X_j não deveria constar do modelo de regressão.

Agresti [Agresti, 2002] dá-nos a estatística de Wald multivariada dada pela expressão

$$W = (\widehat{\beta} - \beta^*)^T \mathfrak{I}(\widehat{\beta})(\widehat{\beta} - \beta^*), \tag{2.51}$$

onde o objetivo é testar as seguintes hipóteses:

$$\text{H0: } \beta = \beta^*$$
$$\text{H1: } \beta \neq \beta^*,$$

considerando, claro está, $\beta = (\beta_0, \beta_1, ..., \beta_p)$, $\widehat{\beta} = (\widehat{\beta_0}, \widehat{\beta_1}, ..., \widehat{\beta_p})$ e $\beta^* = (\beta_0^*, \beta_1^*, ..., \beta_p^*)$.

A estatística anterior segue uma distribuição de χ^2 com $p + 1$ graus de liberdade.

Consideraremos, para um nível de significância α, a hipótese alternativa como verdadeira se conseguirmos rejeitar a hipótese nula, isto é, se verificarmos que $W > \chi^2_{1-\alpha}(p + 1)$.

O teste de score baseia-se na distribuição teórica das derivadas da função de log-verosimilhança. A versão multivariada do teste exige a determinação da matriz das derivadas da função de log-verosimilhança em ordem a cada um dos β.

O teste de score tem como estatística de teste [Dobson, 2002]:

$$ST = U(\beta^*)^T \mathfrak{I}(\widehat{\beta})^{-1} U(\beta^*), \tag{2.52}$$

onde $U(\beta^*)$ e $\mathfrak{I}(\beta^*)$ representam, respetivamente, a estatística score, avaliadas em β^*, e a matriz de informação avaliadas em $\widehat{\beta}$.

A estatística de teste ST segue uma distribuição χ^2 com $p + 1$ graus de liberdade, servindo para avaliar as seguintes hipóteses estatísticas:

$$\text{H0: } \beta = \beta^*$$
$$\text{H1: } \beta \neq \beta^*,$$

onde$\beta = (\beta_0, \beta_1, ..., \beta_p)$, $\widehat{\beta} = (\widehat{\beta_0}, \widehat{\beta_1}, ..., \widehat{\beta_p})$ e $\beta^* = (\beta_0^*, \beta_1^*, ..., \beta_p^*)$.

Quando $n \to +\infty$, isto é, quando o número de observações da base de dados é muito grande, os três testes apresentados são assintóticamente equivalentes [Cox & Hinkley, 1974].

Capítulo 3

Dados com eventos raros

Neste terceiro capítulo introduziremos o conceito de dados com eventos raros. Para uma melhor compreensão deste conceito incluímos exemplos reais de situações onde estes ocorrem. Abordaremos ainda o fenómeno de separabilidade.

Dependendo da área do conhecimento onde um evento raro ocorre este pode ser visto como algo positivo ou negativo. Por este motivo a definição de evento raro vai variando de área para área, muito à mercê da conotação associada à sua ocorrência.

Segundo King e Zeng [King & Zeng, 2001a], dados com eventos raros correspondem a variáveis dependentes binárias que, tendo em conta as suas dimensões, têm dezenas a milhares de vezes mais *não ocorrências do evento de interesse* do que *ocorrências do evento de interesse*, isto é, na variável dependente teremos dezenas a milhares de vezes mais 0's do que 1's, onde 1 denota a *ocorrência do evento de interesse* e 0 a sua *não ocorrência*.

Os mesmos autores fazem referência a vários exemplos de situações nas quais nos podemos deparar com dados com eventos raros. Entre os exemplos apresentados podemos destacar: *guerras, vetos presidenciais, ativismo político e infeções epidemiológicas*. Deixamos de seguidas dois exemplos reais da ocorrência de eventos raros. O primeiro é um exemplo na área da medicina.

Exemplo 3.1 (Medicina) *Uma das áreas do conhecimento onde o tema dos eventos raros mais é estudada é a da medicina. A ocorrência de eventos raros nesta área aparece das mais variadas formas, sendo a mais imediata ao nosso pensamento as doenças raras.*

Por definição, na União Europeia, considera-se uma doença como sendo rara se a sua prevalência é inferior a 5 em cada 10.000 pessoas. Destacam-se assim, doenças tais como: o ***Síndrome de Williams*** [1]*, a* ***Dermatomiosite*** [2] *e o* ***Síndrome de Rubinstein-Taybi*** [3]*.*

[1] O **Síndrome de Williams** carateriza-se, essencialmente, pela falta de 21 genes no cromossoma 7. Devido a esta falta, os portadores do síndrome não são capazes de produzir *elastina*, uma proteína que forma as fibras elásticas – comuns em regiões como a trompa de Eustáquio, a epiglote e a cartilagem da laringe.

[2] A **Dermatomiosite** é uma doença auto-imune classificada como miopatia inflamatória idiopática, ou seja, inflamação das fibras musculares de causa desconhecida. Esta doença caracteriza-se pela fraqueza dos músculos proximais, em especial dos músculos dos ombros, bacia e coxas, bem como por alterações inflamatórias na pele.

[3] O **Síndrome de Rubinstein-Taybi** é um síndrome dismórfico caracterizado por polegares largos, atraso

Contudo, existem outras doenças que não cabem na definição de doenças raras, mas que ainda assim, a sua ocorrência é considerado um evento raro. É aqui que surgem doenças como a ***Tuberculose*** *(que abordaremos mais à frente) ou o* ***Síndrome de Tourette****.*

A título de explicação o **Síndrome de Tourette** é um tipo de desordem de tiques, isto é, de movimentos involuntários repetitivos, bem como de vocalizações repetitivas. Este Síndrome afeta todas as raças, grupos étnicos e idades, mas é 3 a 4 vezes mais comum em meninos do que em meninas. Estimativas indicam que a sua prevalência em crianças entre os 5 e os 18 anos varia entre 0.3 e 0.8% [Oliveira & Massamo, 2012], ou seja, a probabilidade de se verificar aleatoriamente uma criança com síndrome de Tourette é, pelo menos, uma centenas de vezes menor que a probabilidade de uma criança selecionada aleatoriamente da população não sofrer da doença.

O segundo exemplo, apresentado de seguida, corresponde a um exemplo da ocorrência de eventos raros ao nível financeiro.

Exemplo 3.2 (Finanças) *Outra área na qual podemos encontrar várias situações de eventos raros é a área financeira. A este nível são de destacar a ocorrência de incumprimento no pagamento das prestações do crédito à habitação.*

Segundo um artigo do jornal *Público* de 27 de março de 2017 [Crisóstomo & Soares, 2017], no final de 2016 existiam 2.3 *milhões de clientes bancários com empréstimo à habitação*, dos quais o Banco de Portugal estimava que 5.9% tinha crédito vencido, isto é, estavam com incumprimento no pagamento das prestações.

É verdade que o valor de 5.9% de incumprimento no crédito à habitação é bastante superior aos 0.8%, o valor máximo da estimativa para a ocorrência do síndrome de Tourette em crianças em idades entre os 5 e os 18 anos. Ainda assim, note-se que estes 5.9% de incumprimento significam que no universo dos créditos à habitação, cerca de um em cada vinte créditos, correspondem a incumprimento. É verdade que para o setor, este é um valor bastante elevado, mas, ainda assim, dentro da definição anteriormente apresentada para *eventos raros*.

Em título de nota, é também vulgar na literatura sobre este tema ler-se, como definição, que um evento é considerado raro quando o evento de interesse ocorre em menos de 10% das observações. Note-se que esta definição não se afasta muito da anteriormente apresentada.

mental, face peculiar, atraso de crescimento, malformações associadas e uma predisposição para o desenvolvimento de neoplasias.

3.1 Separabilidade

Quando falamos em regressão logística, e principalmente quando a abordamos no domínio dos eventos raros, é fundamental falarmos do fenómeno de **separabilidade**, bem como do problema a ele associado.

Na regressão logística podem surgir situações em que o algoritmo para a determinação da verosimilhança convirja, mas a estimativa de um parâmetro tenda para $\pm\infty$. Normalmente, designamos este fenómeno por **separação**. Vulgarmente, ao nível da literatura atual, é também possível ver este fenómeno a ser referido como "probabilidade monótona".

De um modo geral, o fenómeno de separação está associado ao facto de a ocorrência do evento de interesse e a sua não ocorrência serem facilmente separados por um único fator de risco ou por uma combinação linear, não trivial, de fatores de risco.

A separação ocorre principalmente em amostras pequenas e/ou com várias *variáveis desequilibradas* [Heinze & Schemper, 2002]. Entenda-se aqui variáveis desequilibradas como sendo variáveis categóricas, que quando confrontadas com a variável resposta, apresentam diferenças muito consideráveis entre as várias proporções.

Como nos mostram Heinze e Schemper no artigo *A solution to the problem of separation in logistic regression* de 2002 [Heinze & Schemper, 2002], o problema da separação não é insignificante.

São vários os fatores que podem levar à existência de separabilidade nos dados. Ao nível da literatura encontram-se documentados um conjunto de fatores que conduzem à ocorrência deste fenómeno, dos quais se destacam: o tamanho da amostra, o número de variáveis binárias e a magnitude dos *odds* associados a tais variáveis.

Como é de fácil perceção, num estudo sobre eventos raros, uma das principais causas para o aparecimento do fenómeno de separabilidade está relacionado com o facto de existir um elevado desequilibro nas várias categorias das variáveis. Tais desequilíbrios conduzem ao aparecimento de *odds* com valores indesejados.

No artigo já referenciado, os autores deixam-nos ainda como proposta, para ultrapassar parte dos problemas relacionados com a separabilidade dos dados, a aplicação do modelo de regressão logística de Firth. Este modelo de regressão foi proposto num artigo da autoria de David Firth em 1993. Apresentaremos este modelo mais à frente na presente dissertação, mais propriamente, no Capítulo 6.

Capítulo 4

Regressão logística condicional em estudos de caso-controlo

Neste capítulo são abordados os estudos de caso-controlo, as variáveis de confundimento e o modelo de regressão logística condicional em estudos de caso-controlos emparelhados.

4.1 Introdução

Os estudos de caso-controlo começaram a ser aplicados ao nível da epidemiologia com o objetivo de determinar, por comparação de populações, os fatores de risco que podem estar associam a uma determinada doença.

Nathan Mantel (1919 - 2002) e William Haenszel (1910 - 1998) apresentaram em meados do século XX um artigo sobre os aspetos estatísticos da análise de dados de estudos de caso-controlo, que foram, e ainda são, bastante aplicados tanto por estatísticos como por epidemiologista na busca da cura do cancro e de outras doenças [Breslow & Day, 1980].

Com os avanços científicos após a Segunda Guerra Mundial no campo da tecnologia tornou-se possível e cada vez mais simples realizar uma série de análises exploratórias que antes eram impensáveis [Breslow & Day, 1980].

É neste contexto de expansão tecnologica que surgem os primeiros trabalhos sobre regressão logística condicional em estudos de caso-controlo, apresentados por Norman Breslow (1941 - 2015) e Nicholas Day (1939 - presente) em 1980, no livro *Statistical Methods in Cancer Research, Volume I - The Analysis of Case-Control Studies* [Breslow & Day, 1980].

4.2 Estudos de Caso-Controlo

O primeiro estudo de caso-controlo documentado remonta ao ano de 1926, num estudo sobre o cancro de mama, publicado por Janet Lane-Claypon [Song & Chung, 2010]. Pouco frequentes até à década de 1950, foi nesta altura que foi publicado o famoso estudo que relaciona o tabagismo ao cancro do pulmão, estudo este que acabou por impulsionar a aplicação desta

metodologia de estudo.

Os estudos de caso-controlo são bastante utilizados ao nível das ciências da vida. Estes são bastante utilizados, por exemplo, ao nível da medicina, com o intuito de determinar os fatores de exposição que estão associados a uma determinada doença. Deste modo, com o objetivo de simplificar a abordagem a este tema, suponhamos que estamos num contexto médico, no qual queremos estudar os fatores de exposição que influenciam uma determinada doença.

Neste contexto, o primeiro passo passa pela criação de uma amostra aleatória de indivíduos que sejam portadores da doença. Aos elementos desta amostra, constituída apenas por indivíduos que têm em comum o facto de serem portadores da doença em estudo, designamos por **casos**. O segundo passo diz respeito à recolha de um conjunto de informações sobre os vários casos, de onde se destacam as informações sobre os potenciais fatores de exposição que se pretendem estudar. O terceiro passo corresponde à construção da amostra de **controlos**. Song e Chung [Song & Chung, 2010] dizem-nos que esta pode ser a parte mais exigente de todo o estudo de caso-controlo. Designamos por controlos aos elementos do conjunto de pacientes não portadores da doença que são utilizados para comparar com os casos. É fundamental que estes pacientes tenham uma distribuição sociodemográfica idêntica à dos casos. Por outras palavras, o investigador pode considerar para grupo de controlo uma população em risco, isto é, com o potencial de desenvolver a doença.

De um modo geral, para assegurar a existência de correspondência entre os casos e os controlos é feito um emparelhamento individual entre cada caso e um ou mais controlos. Este emparelhamento é baseado num conjunto de variáveis que designamos por variáveis de confundimento, que são, em geral, variáveis sociodemográficas, mas não só. Abordaremos as variáveis de confundimento na secção seguinte.

Após a amostragem dos controlos é feita a análise estatística com vista à comparação dos fatores de exposição em estudo nas duas populações, de modo a concluir se estes são ou não fatores preponderantes para a prevalência e/ou propagação da doença.

4.3 Variáveis de confundimento

Seguindo o contexto médico, quando estamos a analisar os fatores de exposição que podem influenciar a propagação de uma determinada doença são estudadas uma série de variáveis explicativas, relativas às causas, de modo a perceber qual é o contributo de cada fator na variável resposta. Aqui, entendemos como variável resposta a variável que nos indica sobre a ocorrência ou não da doença.

Entre as variáveis explicativas poderemos encontrar variáveis que representam causas da doença, isto é, que são fatores de exposição, bem como variáveis que distorcem o verdadeiro efeito de um determinado fator sobre a variável resposta [Hosmer & Lemeshow, 2000].

A uma variável explicativa que está associada com a variável respostas e, que ao mesmo tempo, está associada a outra variáveis explicativas designamos por **variável de confundimento**. Num contexto médico, a variável de confundimento pode ser vista como um fator de

risco para a doença em estudo, mas que não é uma causa direta da doença.

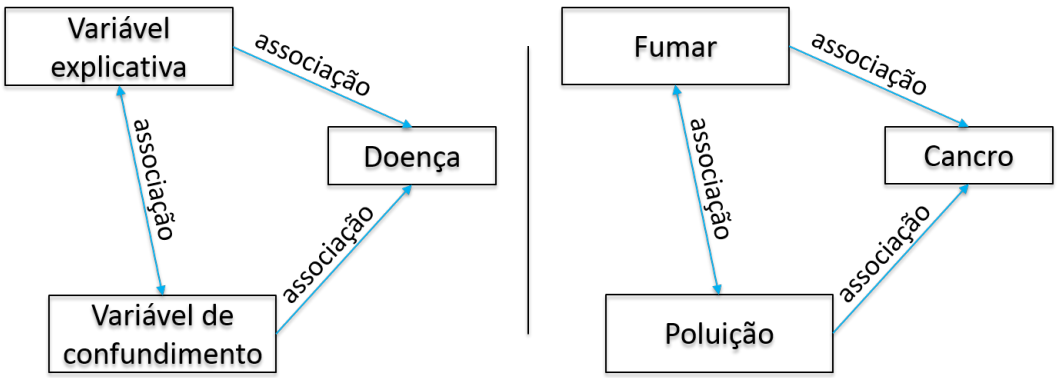


Figura 4.1: Esquerda: esquema da associação da variável de confundimento com a variável resposta e a variável explicativa, num contexto médico. Direita: um exemplo simples.

Geralmente, são consideradas boas variáveis de confundimento o sexo, a idade, o estado matrimonial, a orientação sexual, o grupo étnico, a área de residência, entre várias outras.

Note-se que as variáveis acima indicadas não são, de um modo geral, causas para o aparecimento de doenças, contudo, é fácil de perceber que algumas destas permitem explicar se um certo indivíduo é mais ou menos vulnerável a uma determinada doença.

Como já foi dito anteriormente, as variáveis de confundimento são usadas ao nível dos estudos de caso-controlo de modo a emparelhar os vários casos com um ou mais controlos. Isto é, ao fazermos a amostragem do conjunto dos controlos selecionamos aleatoriamente controlos que compartilham com os casos os mesmos valores/categorias para as variáveis de confundimento [Song & Chung, 2010].

Figura 4.2: Ilustração da correspondência caso-controlo.

Supondo um estudo no qual temos três variáveis de confundimento, as variáveis X_1, X_2 e X_3, cada uma com duas categorias, respetivamente, X_{11} e X_{12}, X_{21} e X_{22} e X_{31} e X_{32}, então

obtemos oito **padrões de confundimento** diferentes. Podemos deste modo dividir os casos e os controlos em oito conjuntos diferentes. Essa divisão encontra-se esquematizada na figura 4.3.

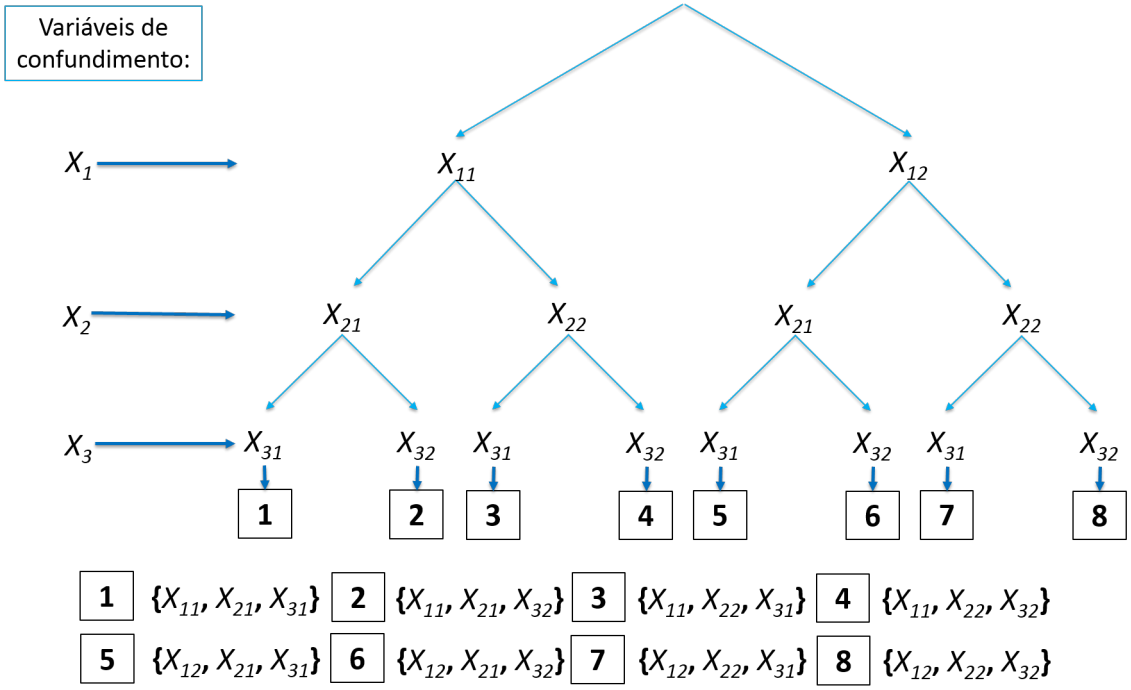


Figura 4.3: Esquema representativo dos possíveis padrões de confundimento que são obtidos com três variáveis de confundimento, cada uma com duas categorias.

4.4 Regressão logística condicional em estudos de caso-controlo emparelhados

Ao nível da literatura é frequente designar-se um estudo para o qual a cada caso se façam corresponder M controlos por **estudo de correspondência 1:M** e ao conjunto formado por cada caso e correspondentes M controlos por **conjunto correspondente**.

David Collett [Collett, 2003] refere que geralmente o número de controlos a corresponderem a cada caso, M, situa-se entre um e cinco. O mesmo autor refere que por vezes, quando o problema em estudo a tal obriga, é apropriado que haja conjuntos correspondentes com mais do que um caso. No entanto, os estudos de correspondência 1:M são os mais amplamente utilizados, sendo nestes que centraremos a nossa atenção.

Assim, consideremos que estamos perante um estudo de correspondência 1:M onde existem n observações que verificam o evento de interesse. Consideremos ainda que a probabilidade de uma observação corresponder a um caso, isto é, corresponder a uma observação que verifica o evento de interesse, depende de p variáveis explicativas.

É de relevo referir que os valores das variáveis de confundimento dentro de cada conjunto correspondente não variam, diferindo apenas entre os n conjunto correspondente.

Denotemos por x_{ij} o vetor contendo a medição das p variáveis explicativa,s $X_1, X2, ..., X_p$, da i-ésima observação do j-ésimo conjunto correspondente, com $i = 0, 1, ..., M$ e $j = 1, 2, ..., n$. Tendo em conta esta notação e para $j = 1, 2, ..., n$, segue que:

- x_{0j} é o vetor que contém o valor das p variáveis explicativas relativas ao caso do j-ésimo conjunto correspondente; e
- $x_{ij}, i = 1, 2, ..., M$, é o vetor que contém o valor das p variáveis explicativas relativas ao i-ésimo controlo do j-ésimo conjunto correspondente.

Denotemos ainda por $\pi_j(x_{ij})$, com $i = 0, 1, ..., M$ e $j = 1, 2, ..., n$, a probabilidade da i-ésima observação do j-ésimo conjunto correspondente verificar o evento de interesse, isto é, $\pi_j(x_{ij}) = E(Y_j|x_{ij})$.

A probabilidade $\pi_j(x_{ij})$ é modelada utilizando o modelo de regressão logística usual, com a diferença de que consideramos um termo constante, α_j, diferente para cada conjunto correspondente. Isto é, para cada $j = 1, 2, ..., n$ temos:

$$logit\{\pi_j(x_{ij})\} = \alpha_j + \beta_1 x_{1ij} + \beta_2 x_{2ij} + ... + \beta_p x_{pij} \tag{4.1}$$

onde x_{kij}, $k = 1, 2, ..., p$, corresponde ao valor da k-ésima variável explicativa, X_k, relativa à i-ésima observação do j-ésimo conjunto correspondente, isto é, $x_{ij}^T = (x_{1ij}, x_{2ij}, ..., x_{pij})$.

Assim, se considerarmos $\beta^T = (\beta_1, \beta_2, ..., \beta_p)$, temos que:

$$logit\{\pi_j(x_{ij})\} = \alpha_j + \beta^T x_{ij}. \tag{4.2}$$

Sendo o $logit\{\pi_j(x_{ij})\} = log\left(\dfrac{\pi_j(x_{ij})}{1-\pi_j(x_{ij})}\right)$ podemos escrever a expressão (4.2) da seguinte forma:

$$\pi_j(x_{ij}) = \frac{e^{\alpha_j + \beta^T x_{ij}}}{1 + e^{\alpha_j + \beta^T x_{ij}}}, \quad j = 1, 2, ..., n. \tag{4.3}$$

Chegados a estes ponto é necessário construir uma função de verosimilhança condicionada, de modo a que as correspondências entre os casos e os controlos sejam utilizadas no ajustamento do modelo de regressão. Tal função de verosimilhança corresponde ao produto de n termos, cada um dos quais referente à probabilidade condicionada do caso em cada conjunto correpondente [Collett, 2003].

Segundo D. C. Thomas [Liddell & et al., 1977] a função de verosimilhança condicionada para um estudo de correspondência 1:M é dada pela expressão

$$\begin{aligned} L(\beta) &= \prod_{j=1}^{n} \frac{e^{\beta^T x_{0j}}}{\sum_{i=0}^{M} e^{\beta^T x_{ij}}} \\ &= \prod_{j=1}^{n} \frac{exp(\sum_{k=1}^{p} \beta_k x_{0jk})}{\sum_{i=0}^{M} exp(\sum_{k=1}^{p} \beta_k x_{ijk})}. \end{aligned} \tag{4.4}$$

Vamos de seguida verificar a expressão (4.4). Para tal tenha-se em conta que $P(x_{ij}|Y=1)$ é a probabilidade de uma observação que verifica o evento de interesse no j-ésimo conjunto ter variáveis explicativas $x_{ij},\ i=0,1,...,M$, e que $P(x_{ij}|\overline{Y=1})$ é a probabilidade de uma observação que não verifica o evento de interesse ter variáveis explicativas x_{ij}. Assim, se x_{0j} corresponder às medições das p variáveis explicativas de um caso e se $x_{ij}, i \geqslant 1$, corresponder às medições das p variáveis explicatidas dos controlos, então a probabilidade conjunta de x_{0j} corresponder ao caso e de x_{ij} aos controlos é

$$P(x_{0j}|Y=1)\prod_{i=1}^{M} P(x_{ij}|\overline{Y=1}).$$

Note-se ainda que a probabilidade de uma das $M+1$ observações do j-ésimo conjunto correspondente ser um caso e as restantes serem controlos é dada pela união entre a probabilidade da observação $i=0$ corresponder ao caso e as restantes aos controlos, com a probabilidade da observação $i=1$ corresponder ao caso e as restantes correspondem aos controlos, e assim sucessivamente. Sendo $I=\{0,1,...,M\}$, podemos representar esta probabilidade utilizando a seguinte expressão:

$$\sum_{i\in I} P(x_{ij}|Y=1)\prod_{r\in I,\ r\neq i} P(x_{rj}|\overline{Y=1}).$$

Assim, num estudo de correspondência $1:M$, a probabilidade condicionada de no j-ésimo conjunto correspondente a observação $i=0$ corresponder ao caso e as restantes M observações corresponderem aos controlos, é dada pelo rácio entre as duas últimas expressões, ou seja, é da forma:

$$\frac{P(x_{0j}|Y=1)\prod_{i=1}^{M} P(x_{ij}|\overline{Y=1})}{\sum_{i\in I} P(x_{ij}|Y=1)\prod_{r\in I,\ r\neq i} P(x_{rj}|\overline{Y=1})}. \tag{4.5}$$

De acordo com o **Teorema de Bayes**, a expressão (4.5) pode ser expressa da forma:

$$\frac{P(Y=1|x_{0j})\prod_{i=1}^{M} P(\overline{Y=1}|x_{ij})}{\sum_{i\in I} P(Y=1|x_{ij})\prod_{r\in I,\ r\neq i} P(\overline{Y=1}|x_{rj})},$$

o que pode ser reduzido à seguinte expressão:

$$\begin{aligned}&\left\{1+\frac{\sum_{i=1}^{M} P(Y=1|x_{ij})\prod_{r\in I,\ r\neq i} P(\overline{Y=1}|x_{rj})}{P(Y=1|x_{0j})\prod_{i=1}^{M} P(\overline{Y=1}|x_{ij})}\right\}^{-1}\\ &=\left\{1+\sum_{i=1}^{M}\frac{P(Y=1|x_{ij})P(\overline{Y=1}|x_{0j})}{P(Y=1|x_{0j})P(\overline{Y=1}|x_{ij})}\right\}^{-1}\end{aligned}$$

Note-se que $P(Y=1|x_{ij})=\pi_j(x_{ij})=\dfrac{e^{\alpha_j+\beta^T x_{ij}}}{1+e^{\alpha_j+\beta^T x_{ij}}}$, pelo que a expressão anterior pode ser reescrita como:

$$\left\{1+\sum_{i=1}^{M}\frac{e^{\alpha_j+\beta^T x_{ij}}}{e^{\alpha_j+\beta^T x_{0j}}}\right\}^{-1} \tag{4.6}$$

Simplificando esta expressão:

$$\left\{1+\sum_{i=1}^{M}\frac{e^{\alpha_j+\beta^T x_{ij}}}{e^{\alpha_j+\beta^T x_{0j}}}\right\}^{-1}=\left\{\sum_{i=0}^{M}\frac{e^{\beta^T x_{ij}}}{e^{\beta^T x_{0j}}}\right\}^{-1}=\frac{e^{\beta^T x_{0j}}}{\sum_{i=0}^{M}e^{\beta^T x_{ij}}} \tag{4.7}$$

Assim a função de verosimilhança condicional para todos os n conjuntos correspondentes é dada pela expressão:

$$L(\beta)=\prod_{j=1}^{n}\frac{e^{\beta^T x_{0j}}}{\sum_{i=0}^{M}e^{\beta^T x_{ij}}}$$

Fica assim verificada a expressão (4.4) de D. C. Thomas.

Podemos simplificar a expressão (4.4) da seguinte forma:

$$L(\beta)=\prod_{j=1}^{n}\frac{e^{\beta^T x_{0j}}}{\sum_{i=0}^{M}e^{\beta^T x_{ij}}}=\prod_{j=1}^{n}\frac{1}{\sum_{i=0}^{M}e^{\beta^T (x_{ij}-x_{0j})}}=\prod_{j=1}^{n}\left\{1+\sum_{i=1}^{M}e^{\beta^T (x_{ij}-x_{0j})}\right\}^{-1} \tag{4.8}$$

Com esta simplificação obtemos a expressão apresentada por Breslow e Day [Breslow & Day, 1980]:

$$L(\beta)=\prod_{j=1}^{n}\left\{1+\sum_{i=1}^{M}exp\left[\sum_{k=1}^{p}\beta_k(x_{ijk}-x_{0jk})\right]\right\}^{-1}, \tag{4.9}$$

onde $\beta^T=(\beta_1,\beta_2,...,\beta_p)$.

É com base nesta função de verosimilhança que se determinam os estimadores de máxima verosimilhança para os β's.

Collett [Collett, 2003] diz-nos que para a determinação dos parâmetros do modelo de regressão logística para estudos de correspondência caso-controlo usa-se o mesmo método que para os modelos lineares generalizados. O mesmo autor mostra que a função de verosimilhança em estudo é equivalente à função de verosimilhança do modelo de regressão de Poisson, com função de ligação logarítmica.

Ao nível do R: Para a obtenção do modelo de regressão logística condicional em estudos de caso-controlo emparelhados no R podemos utilizar a função **clogit( )** pertencente à biblioteca **survival**. Esta função é usada para a obtenção de vários modelos de regressão logística condicional, sendo essencial definir o parâmetro "*method="exact""*para obtermos dos modelos de regressão que pretendemos. A função apresenta ainda vários outros parâmetros. Apresenta-se de seguida a função e os respetivos parâmetros:

```
1 clogit(case.status~exposure+strata(matched.set), data, weights,
2        subset, na.action, method="exact", ...)
```

Mais informações sobre esta função podem ser encontradas no manual do R, em `https://stat.ethz.ch/R-manual/R-devel/library/survival/html/clogit.html`.

Capítulo 5

Regressão logística com correção do viés, usando correção apriori e pesos

Neste capítulo seguiremos a abordagem proposta por King e Zeng em 2001.

5.1 Introdução

Nos artigos *Logistic regression in rare events data* [King & Zeng, 2001a] e *Explaining rare events in international relations* [King & Zeng, 2001b] os autores Gary King e Langche Zeng exploram a ocorrência de eventos raros nas ciências políticas, deixando-nos duas propostas de como corrigir o viés nos estimadores de máxima verosimilhança dos parâmetros do modelo de regressão logística. Ao mesmo tempo, estes autores, abordam o tema da amostragem do conjunto de treino para o ajustamento do modelo, utilizando a correspondência de caso-controlo.

As duas propostas de King e Zeng, e que apresentamos nas secções de seguida, são a correção apriori e a correção ponderada (utilizando pesos), que visam a correção do viés gerado pela amostragem no contexto de caso-controlo e, ao mesmo tempo, dos casos corresponderem a eventos raros.

O aparecimento da correção apriori proposta por King e Zeng é atribuído a vários autores. Epidemiologistas e bioestatísticos geralmente atribuem o seu surgimento a Prentice e Pyke (1979); já os economistas atribuem, de um modo geral, o seu aparecimento a Manski e Lerman (1977).

5.2 Regressão logística com correção apriori

A correção apriori é o método mais simples de corrigir o viés de uma regressão logística no contexto de uma amostra de casos e controlos emparelhados [King & Zeng, 2001a]. O procedimento corresponde a ajustar o modelo de regressão logística usual e corrigir os estimadores, tendo por base informação externa sobre a proporção de casos na população e a proporção de

casos no conjunto de treino. De um modo geral, a proporção de casos na população é denotada por τ e a proporção de casos na amostra que serve de treino para o ajustamento do modelo por $\overline{y}$.

Para o modelo *logit* o estimador de máxima verosimilhança $\widehat{\beta}_1$, da equação $logit\{\pi(x)\} = \beta_0+\beta_1 x_1+...+\beta_p x_p$, é um estimador estatisticamente consistente de β_1. Deste modo a correção do estimador para este parâmetro é desnecessária. De um modo mais genérico, os coeficientes de declive (aqueles que refletem os efeitos das variáveis explicativas) são consistentes, pelo que o interesse da correção recai apenas sobre o estimador de máxima verosimilhança do parâmetro β_0, isto é, de $\widehat{\beta}_0$.

A correção que King e Zeng propuseram, em 2001, é a seguinte:

$$\widehat{\beta}_0 - ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\overline{y}}{1-\overline{y}}\right)\right]. \tag{5.1}$$

Uma das vantagens fundamentais da correção apriori é a sua fácil utilização. Note-se que esta pode ser utilizada em qualquer software de estatística que possa ser empregado na estimação de coeficientes do modelos *logit*, uma vez que a expressão anterior é facilmente aplicada ao *termo constante*.

A única desvantagem da aplicação da correção apriori é que se o modelo não for bem especificado as estimativas dos parâmetros de regressão são ligeiramente menos robustas do que quando é usada regressão logística ponderada.

5.3 Regressão logística ponderada

O outro método proposto pelos autores já mencionados para a correção do viés é a ponderação dos dados através de pesos. Esta ponderação visa compensar as diferenças existentes entre a amostra e a população induzidas pela amostragem.

O estimador de máxima verosimilhança resultante deste método de ponderação da regressão logística, o qual geralmente designamos por estimador de máxima verosimilhança exógeno, deve o seu aparecimento a um trabalho denominado por *The Estimation of Choice Probabilities from Choice Based Samples*, do ano de 1977, da autoria de Charles Manski e Steven Leman. Este estimador é relativamente simples, tal como poderemos ver de seguida.

Nesta situação em vez de querermos maximizar a função de log-verosimilhança usual, queremos maximizar a função de log-verosimilhança ponderada:

$$\begin{aligned} log(L_\omega(\beta|y)) &= \omega_1 \sum_{Y_i=1} log(\pi_i) + \omega_0 \sum_{Y_i=0} log(1-\pi_i) \\ &= -\sum_{i=1}^{n} \omega_i \; log(1+e^{(1-2y_i)x_i\beta}), \end{aligned} \tag{5.2}$$

sendo os pesos $\omega_1 = \tau/\overline{y}$ e $\omega_0 = (1-\tau)/(1-\overline{y})$, onde

$$\omega_i = \omega_1 Y_i + \omega_0(1-Y_i).$$

Segundo King e Zeng [King & Zeng, 2001a] a correção utilizando a ponderação pode ser mais precisa que a correção apriori quando a amostra é grande. Contudo, esta é assintoticamente menos eficiente que a correção apriori, o que pode ser avaliado com amostras mais pequenas, não sendo, ainda assim, as diferenças muito significativas.

5.4 Estimação dos parâmetros - eventos raros

Como parece ser fácil de antever, numa situação de eventos raros, a probabilidade de ocorrer o evento de interesse, isto é $P(Y = 1)$, é subestimada, sendo a probabilidade de este não ocorrer, $P(Y = 0)$, subrestimada.

Para podermos intuitivamente observar o acima descrito, podemos considerar um caso simples, no qual a variável resposta depende apenas de uma variável explicativa.

Atente-se na figura 5.1:

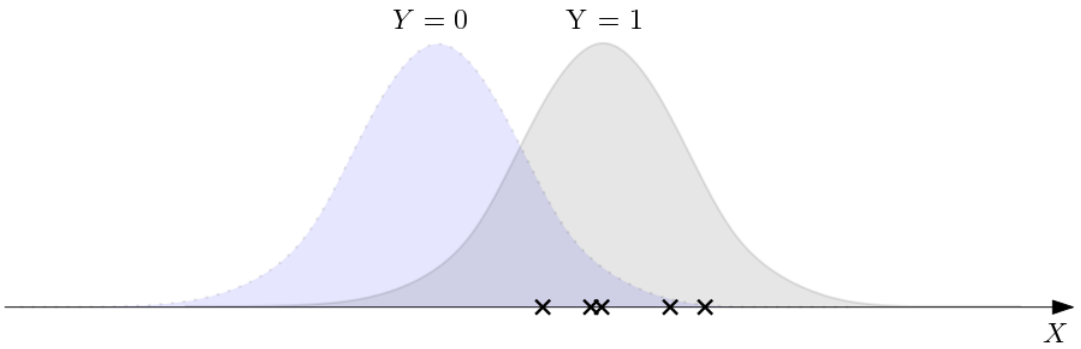


Figura 5.1: Ilustração das funções de densidade de $X|Y = 0$ e de $X|Y = 1$.

Na figura 5.1 podemos ver uma situação de eventos raros onde o número de observações para as quais não ocorre o evento de interesse ($Y = 0$) é muito superior ao número de observações nas quais se verifica a ocorrência ($Y = 1$). Na ilustração, podemos ver a representação das funções de densidade de $X|Y = 0$ (mais à esquerda) e de $X|Y = 1$ (mais à direita), sendo, no eixo das abcissas representados por uma cruz os valores de $X|Y = 1$.

Dado o volume de observações, conseguimos estimar quase sem erros [1] a variável resposta para as observações com $Y = 0$. Em contraponto, como existem poucas observações com $Y = 1$, observa-se uma estimação com alguns erros [2]. O reduzido número de observações enfraquece a determinação da função de densidade de $X|Y = 1$, gerando ainda tendências em direção às caudas que são muito curtas.

Assim, e tendo por objetivo a supressão do viés gerado pelos eventos raros, as estimativas dos coeficientes do modelo de regressão são corrigidas.

[1] No exemplo construido em R para avaliar o que era descrito pela literatura, o erro obtido ao nível das observações com $Y = 0$ foi pouco superior a 1%.

[2] O que foi possível confirmar com o exemplo criado em R.

McCullagh e Nelder [McCullagh & Nelder, 1989] mostraram que o viés é dado pela seguinte expressão:

$$\text{viés}(\widehat{\beta}) = (X^T W X)^{-1} X^T W \xi \tag{5.3}$$

onde $\xi_i = 0.5 Q_{ii}[(1+\omega_1)\widehat{\pi}_i - \omega_1]$, sendo Q_{ii} os elementos da diagonal de $Q = X(X^T W X)^{-1} X^T$ e $W = diag[\widehat{\pi}_i(1 - \widehat{\pi}_i \omega_i]$.

Deste modo, obtêm-se os estimadores de máxima verosimilhança corrigidos fazendo:

$$\beta^* = \widehat{\beta} - \text{viés}(\widehat{\beta}). \tag{5.4}$$

Ao nível do R: Para a obtenção do modelo de regressão logística com as correções propostas por King e Zeng no R podemos utilizar a função **relogit( )** pertencente à biblioteca **Zelig**. Esta função é usada para a obtenção de vários modelos de regressão logística com correção, sendo essencial definir o parâmetro "*tau*"com a proporção de casos na população, de modo a obtermos os modelos de regressão que pretendemos. A função apresenta ainda vários outros parâmetros. Apresenta-se de seguida a função e os respetivos parâmetros:

```
1 relogit(formula, tau = NULL, case.correct = c("prior", "weighting"),
2         bias.correct = TRUE, robust = FALSE, data = mydata, ...))
```

Mais informações sobre esta função podem ser encontradas no manual do R, em `http://ftp.uni-bayreuth.de/math/statlib/R/CRAN/doc/vignettes/Zelig/relogit.pdf`.

Capítulo 6

Regressão logística de Firth

Neste capítulo abordaremos o modelo de redução do viés dos estimadores de máxima verosimilhança, proposto por David Firth em 1993 [Firth, 1993].

6.1 Introdução

A redução/correção do viés na estimação de máxima verosimilhança dos parâmetros da regressão logística tem sido estudada por muitos autores nas últimas décadas. Destacam-se autores como: Anderson & Richardson (1979), McLachlan (1980), Schaefer (1983), Copas (1988), McCullagh & Nelder (1989) e Cordeiro & McCullagh (1991) [Firth, 1993].

Num problema em que o parâmetro a estimar utilizando o estimadores de máxima verosimilhança é um vetor da forma $\theta = (\theta_1, \theta_2, ..., \theta_p)$, o viés assimptótico do estimador de máxima verosimilhança, $\widehat{\theta} = (\widehat{\theta}_1, \widehat{\theta}_2, ..., \widehat{\theta}_p)$, pode ser escrito da seguinte forma:

$$b(\theta) = \frac{b_1(\theta)}{n} + \frac{b_2(\theta)}{n^2} + ... = \sum_{i=1}^{\infty} \frac{b_i(\theta)}{n^i}, \tag{6.1}$$

sendo n o número de observações [Firth, 1993].

Na secção que se segue falaremos sobre um método de reduzir o viés, sendo o objetivo principal a remoção do termo $O(n^{-1})$.

O que levou David Firth a desenvolver a redução de viés como ela será apresentada de seguida deve-se ao facto de à época, os métodos que eram usados para a correção serem, nas suas palavras, apenas "corretivos"e pouco "preventivos", uma vez que os estimadores de máxima verosimilhança eram primeiro calculados e só depois corrigidos.

Será abordada, ao longo deste capítulo, a ideia global que está por detrás da redução do viés, modificando a função dos scores.

6.2 Redução do viés modificando a função dos scores

O que abordamos de seguida não se restringe apenas à regressão logística. Apresenta-se a ideia global por detrás da redução do viés, usando uma modificação na função dos scores.

Já vimos nesta dissertação que para obtermos o estimador de maxima verosimilhança de um determinado parâmetro consideramos a solução da seguinte equação:

$$U(\theta) = 0,$$

onde $U(\theta)$ corresponde à derivada da função de log-verosimilhança em ordem a θ.

A ideia que está por trás do trabalho de Firth é o de reduzir o viés no estimador de máxima verosimilhança, $\widehat{\theta}$, induzindo um pequeno viés na função dos scores.

Geometricamente, esta correção corresponde ao que podemos observar na figura (6.1).

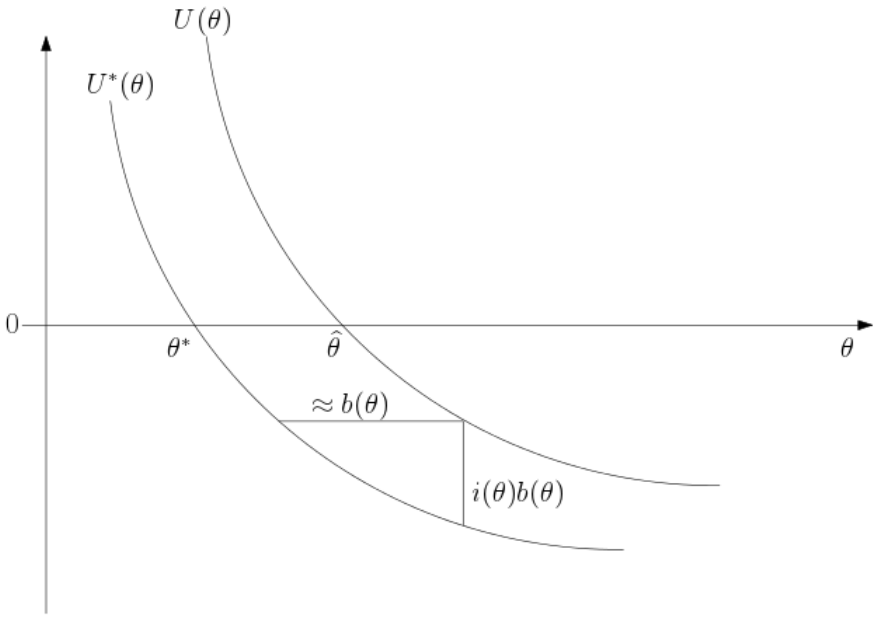


Figura 6.1: Correção do enviesamento na função dos scores (baseado em [Firth, 1993]).

Se $\widehat{\theta}$ está sujeito a um viés positivo, $b(\theta)$, então pode ser considerado um viés negativo que tente corrigir o viés no estimador de máxima verosimilhança. Assim, pode em cada ponto θ ser considerado um valor $i(\theta)b(\theta)$, onde $-i(\theta) = U'(\theta)$ é o gradiente local.

Deste modo, a função dos scores modificado é da forma:

$$U^*(\theta) = U(\theta) - i(\theta)b(\theta). \tag{6.2}$$

Assim, existe θ^*, tal que $U^*(\theta^*) = 0$, o que designamos por estimador de máxima verosimilhança modificado.

Antes de formalizarmos argumentos para a solução do nosso problema, vamos empregar uma notação para as derivadas da função de log-verosimilhança e os seus cumulantes conjuntos nulos.

Assim, seguindo uma notação idêntica à utilizada por McCullgh [McCullagh, 1987], consideremos:

$$U_r(\theta) = \frac{\partial l}{\partial \theta_r} \quad e \quad U_{rs}(\theta) = \frac{\partial^2 l}{\partial \theta_r \theta_s}, \tag{6.3}$$

onde $\theta = (\theta_1, \theta_2, ..., \theta_p)$ é um parâmetro vetor.

Os cumulantes conjuntos nulos são dados por:

$$K_{r,s} = n^{-1}E\{U_r U_s\}, \quad K_{r,s,t} = n^{-1}E\{U_r U_s U_t\}, \quad K_{r,st} = n^{-1}E\{U_r U_{st}\} \tag{6.4}$$

e assim sucessivamente. Lembremos ainda que relativamente aos cumulantes conjuntos nulos, se verificam as seguintes igualdades:

$$K_{rs} + K_{r,s} = 0 \quad e \quad K_{rst} + K_{r,st} + K_{s,rt} + K_{t,rs} + K_{r,s,t} = 0. \tag{6.5}$$

Consideremos agora uma modificação bastante geral da função dos scores da forma

$$U_r^*(\theta) = U_r(\theta) + A_r(\theta), \tag{6.6}$$

onde $A_r(\theta)$ está sempre dependente dos dados e é $O_p(1)$ quando $n \to \infty$. Deste modo, se supusermos que $\widehat{\theta}$ e θ^* são tais que satisfazem, respetivamente, $U(\widehat{\theta}) = 0$ e $U^*(\theta^*) = 0$. Escrevendo $\widehat{\gamma} = n^{\frac{1}{2}}(\theta^* - \theta)$ e usando, à semelhança de Firth [Firth, 1993], um argumento fechado utilizado por McCullag [McCullagh, 1987], baseado na expansão de $U_r^*(\theta^*)$ sobre o valor de θ, o viés de θ^*, temos que:

$$E(\theta^* - \theta)^r = n^{-1}K^{r,s}\{-K^{t,u}(K_{s,t,u} + K_{s,tu})/2 + \alpha_s\} + O(n^{-\frac{3}{2}}), \tag{6.7}$$

onde $K^{r,s}$ denota a inversa da matriz de informação de Fisher $K_{r,s}$ e α_s denota a expectativa nula de A_s.

Na expressão anterior, o termo

$$-n^{-1}K^{r,s}K^{t,u}(K_{s,t,u} + K_{s,tu})/2 = n^{-1}b_1^r(\theta), \tag{6.8}$$

ou seja, corresponde ao viés de primeira ordem de $\widehat{\theta}$ [Firth, 1993].

Assim, a utilização de A_r serve para a remoção do termo de primeira ordem do viés, caso se verifique que

$$K^{r,s}\alpha_s = -b_1^r + O(n^{-\frac{1}{2}}), \tag{6.9}$$

ou, de forma equivalente,

$$\alpha_r = -K_{r,s}b_1^s + O(n^{-\frac{1}{2}}). \tag{6.10}$$

Numa notação matricial, o vetor A deve ser tal que:

$$E(A) = -i(\theta)b_1(\theta)/n + O(n^{-\frac{1}{2}}) \tag{6.11}$$

Tendo em conta esta conclusão, existem dois candidatos mais ou menos óbvios para a escolha de A. À semelhança de Firth (1993), iremos denomina-los por $A^{(E)}$ e $A^{(O)}$, onde:

$$A^{(E)} = -i(\theta)b_1(\theta)/n \quad e \quad A^{(O)} = -I(\theta)b_1(\theta)/n, \tag{6.12}$$

usando, respetivamente, a informação esperada e a informação observada.

No caso de a distribuição envolvida pertencer à família exponencial, então a informação observada $I(\theta)$ coincide com a informação esperada $i(\theta)$, pelo que, $A^{(E)}$ e $A^{(O)}$ coincidem.

6.3 Redução do viés utilizando a distribuição apriori de *Jeffreys*

A utilização da distribuição apriori de *Jeffreys* como função de penalização constitui uma restrição do exposto anteriormente à família exponencial.

Se θ é um parâmetro pertencente a um modelo da família exponencial, então temos que $K_{r,st} = 0$, para todo o r, s e t. Deste modo, o r-ésimo elemento de $A^{(E)}(\theta)$, ou equivalentemente $A^{(O)}(\theta)$, é, considerando (6.5), dado por:

$$a_r = -nK_{r,s}b_1^s/n = K_{r,s}K^{r,t}K^{u,v}K_{t,u,v}/2 = K^{u,v}K_{r,u,v}/2 = -K^{u,v}K_{ruv}/2. \tag{6.13}$$

Utilizando uma notação matricial, podemos escrever a expressão anterior da seguinte forma:

$$a_r = \frac{1}{2}\ tr\left\{i^{-1}\left(\frac{\partial i}{\partial \theta_r}\right)\right\} = \frac{\partial}{\partial \theta_r}\left\{\frac{1}{2}log \mid i(\theta) \mid\right\}. \tag{6.14}$$

A solução de $U_r^* = U_r + a_r = 0$ localiza-se num ponto estacionário de

$$l^*(\theta) = l(\theta) + \frac{1}{2}log \mid i(\theta) \mid, \tag{6.15}$$

ou equivalentemente, da função de verosimilhança penalizada:

$$L^*(\theta) = L(\theta) \mid i(\theta) \mid^{\frac{1}{2}}. \tag{6.16}$$

A função de penalização concluída é $\mid i(\theta) \mid^{\frac{1}{2}}$, que corresponde à expressão da distribuição apriori de *Jeffreys* (que é invariante por reparametrizações do parâmetro θ), introduzida por H. Jeffreys em 1946.

Neste caso, a remoção do termo de primeira ordem do viés é calculado posteriormente, baseando-se neste cálculo apriori [Firth, 1993].

6.4 Redução do enviesamento em regressão logística

Comecemos por considerar que a probabilidade de a i-ésima observação corresponder a um sucesso é $\pi_i = exp(\eta_i)/\{1 + exp(\eta_i)\}$, sendo $\eta_i = \sum x_{ir}\beta_r$.

Ao nível da regressão logística a matriz de informação de Fisher é dada pela expressão $\mathfrak{I}(\theta) = X^T W X$, onde $X = [x_{ir}]$ e $W = diag\{m_i\pi_i(1-\pi_i)\}$, sendo m_i o índice binomial para a i-ésima contagem.

O termo de primeira ordem do viés, ou seja, $O(n^{-1})$ é dado por [McCullagh & Nelder, 1989]:

$$b_1/n = (X^T W X)^{-1} X^T W \xi \tag{6.17}$$

onde $W\xi$ tem como i-ésimo elemento $h_i(\pi - \frac{1}{2})$, sendo h_i o i-ésimo elemento da diagonal da matriz

$$H = W^{\frac{1}{2}} X (X^T W X)^{-1} X^T W^{\frac{1}{2}}. \tag{6.18}$$

A ideia de Firth passa por substituir a equação dos scores $U_r = 0$ pela equação dos scores modificados $U_r^* = 0$.

Note-se que U_r^* é a r-ésima componente de $U^* = U - X^T W \xi$. Em particular,

$$U_r^* = \sum_i \{(y_i + h_i/2) - (m_i + h_i)\pi_i\} x_{ir} \quad (r = 1, ..., p). \tag{6.19}$$

Chegados a este ponto, os estimadores de máxima verosimilhança de β podem ser obtidos iterativamente, utilizando por exemplo o método iterativo dos mínimos quadrados ponderados, já abordados nesta dissertação.

A $(s+1)$-ésima aproximação no método iterativo a dada pela expressão

$$\beta^{(s+1)} = \beta^{(s)} + \mathfrak{I}^{-1}(\beta^{(s)}) U^*(\beta^{(s)}) \tag{6.20}$$

onde $\beta^{(s)}$ é a aproximação obtida na s-ésima iteração.

Alternativamente ao método apresentado acima, os estimadores da regressão logística de Firth podem ser obtidos pela divisão de cada observação i em duas novas observações, tendo valores de resposta y_i e $1 - y_i$ com atualização iterativa de pesos $1 + h_i/2$ e $h_i/2$, respetivamente [Heinze & Schemper, 2002].

A contribuição de uma nova observação para a função de score é então

$$\{(y_i - \pi_i)(1 + h_i/2) + (1 - y_i - \pi_i)h_i/2\} x_{ir} = \{y_i - \pi_i + h_i(1/2 - \pi_i)\} x_{ir}. \tag{6.21}$$

A divisão de cada observação em duas, como descrito anteriormente garante-nos que as estimativas obtidas são finitas.

Deste modo, utilizando a regressão logística de Firth, o problema da separabilidade fica solucionado.

Ao nível do R: Para a obtenção do modelo de regressão logística de Firth no R podemos utilizar a função **brglm( )** pertencente à biblioteca **brglm**. Para a obtenção do modelo de regressão pretendido, isto é, com a penalização do viés na função de verosimilhança é essencial ter o parâmetro "*family = binomial*". A função apresenta ainda vários outros parâmetros. Apresenta-se de seguida a função e os respetivos parâmetros:

```
brglm(formula, family = binomial, data, weights, subset, na.action, offset,
      start = NULL, etastart, mustart, control.glm = glm.control1(...),
      model = TRUE, method = "brglm.fit", pl = FALSE, x = FALSE, y = TRUE,
      contrasts = NULL, control.brglm = brglm.control(...), ...)
```

Mais informações sobre esta função podem ser encontradas no manual do R, em `https://cran.r-project.org/web/packages/brglm/brglm.pdf`.

Capítulo 7

Aplicação a dados reais

Neste capítulo será feita a aplicação dos vários modelos anteriormente descritos a um conjunto de dados sobre a Tuberculose em Portugal.

7.1 Introdução

Os autores de *Introdução à estatística* [Murteira & et al., 2007] começam um dos seus capítulos com a seguinte frase: *"A estatística tem-se difundido de forma tão rápida, que se torna cada vez mais difícil listar os ramos da atividade humana em que a sua aplicação se tem revelado fecunda ou mesmo indispensável"*. Esta frase traduz o inegável papel que a estatística ocupa no avanço da ciência e, inevitavelmente, do conhecimento. Para corroborar esta ideia, podemos ter em conta o levantamento feito em 1988 por Erich L. Lehman [Lehmann, 1988] no qual este nos indica quarenta e sete áreas onde a estatística possui aplicações. Entre estas áreas podemos ler: "*Crystallography*", "*Econometrics*", "*Finance*", "*Fisheries research*", "*Geology*", "*Linguistics*", "*Sociology*", "*Law*"e "*Epidemiology*". É nesta última, epidemiologia [1] que faremos a aplicação que apresentamos no presente capítulo.

A epidemiologia é definida como sendo o estudo da distribuição e dos determinantes de eventos relacionados à saúde em populações específicas, e a aplicação deste estudo à prevenção e controle de problemas de saúde. Esta não se preocupa apenas com a morte, a doença e a deficiência, mas também com os estados de saúde mais positivos e, mais importante, com os meios para melhorar a saúde [Bonita & et al., 2006].

Ao longo das últimas décadas foram sendo criados vários *softwares* para o manuseamento de dados e para a obtenção de conhecimento a partir destes. De entre todos os *softwares* criados destacaria: **SPSS**, **SAS**, **Stata**, **MATLAB**, **Python** e **R**.

A aplicação que apresentamos no presente capítulo será feita em **R**.

[1] A palavra **epidemiologia** tem origem em três palavras da língua grega - "*epi*", "*demos*"e "*logos*", que significam, respetivamente, "*sobre*", "*pessoas*"e "*estudo*".

7.2 Um pouco sobre o software R

O **R** é uma linguagem e ambiente para análises estatísticas e gráficas. Este apareceu por volta do ano de 1993, como resultado de um esforço colaborativo de pessoas de todo o mundo. O *software* foi escrito originalmente por *Robert Gentleman* e *Ross Ihaka* - conhecidos como "**R** & **R**", do Departamento de Estatística da Universidade de Auckland - Nova Zelândia.

A linguagem **R** é um projeto **GNU** que é semelhante à linguagem **S**. O ambiente foi desenvolvido na Bell Laboratories por *John Chambers* e pelos seus colegas.

O **R** está disponível como *software* livre sob os termos da *Free Software Foundation's GNU General Public License* em forma de código fonte [R Project, 2017]. Encontra-se projetado em torno de uma verdadeira linguagem de computador, o que permite aos usuários adicionarem novas funcionalidades adicionais.

⋆ ⋆ ⋆

A versão do **R** utilizada nesta dissertação é a **3.3.3** de 06 de março de 2017. O erro da máquina associado ao uso do R é de 1.110223×10^{-16}. Este valor foi obtido utilizando o seguinte código:

```
1 eps <- 1.0; while(eps + 1.0 > 1.0){eps = eps/2}; print(eps)
```

7.3 Conjunto de dados

O conjunto de dados sobre os quais recairá a análise apresentada ao longo deste capítulo consiste de dados sobre a Tuberculose em Portugal, recolhidos entre os anos de 2008 e 2015.

7.3.1 A Tuberculose

A **tuberculose** (**TB**) é uma infeção causada por um microrganismo chamado *Microbacterium tuberculosis*, também conhecido por bacilo de Koch [Lança, 2003]. Estes microrganismos (bactérias) atacam, de um modo geral, os pulmões, mas podem atacar qualquer parte do corpo [CDC, 2017].

Esta doença ocorre um pouco por todo o mundo, tendo a sua incidência aumentado na década de 80, com o surgimento do Síndrome da Imunodeficiência Adquirida (SIDA).

A sua prevalência é superior em áreas do mundo onde se verifica maior pobreza, desnutrição, falta de acompanhamento médico e má higiene. A doença tem elevada incidência em grupos de risco bem identificados, como são o caso dos sem-abrigo, dos dependentes de drogas injetáveis, dos seropositivos e dos presos.

A **TB** é uma doença que se transmite pelo ar contaminado. Ao tossir, respirar e falar, o doente infetado espalha as gotículas contaminadas no ar, que podem sobreviver por várias horas, desde que não tenham contacto com a luz solar.

Uma pessoa sã ao respirar num ambiente contaminado acaba por inalar gotículas dispersas no ar. Esta inalação faz com que a bactéria se deposite nos pulmões [FPP, 2017]. Nesta altura pode ocorrer uma de três situações. (1) O sistema imunitário consegue eliminar os bacilos e o indivíduo permanece saudável; (2) os bacilos vencem as defesas do organismo evoluindo para doença com o aparecimento de alguns sintomas, tais como tosse, expetoração prolongada, emagrecimento e suor excessivo à noite [Lança, 2003]; ou (3) o sistema imunitário não consegue eliminar eficazmente o bacilo, mas consegue mantê-lo inativo no interior do organismo. Esta situação pode durar por anos, ou mesmo o resto da vida. Estes indivíduos mantém-se saudáveis, não estando doentes, nem contagiando outras pessoas, ainda assim apresentam a possibilidade de vir a ficar doentes, num momento em que o sistema imunitário se mostre mais frágil.

O tratamento da **TB** dura (em geral) seis meses, podendo durar mais tempo, dependendo da resposta do organismo ao tratamento. Este é feito com base em medicamentos designados por anti-bacilares, administrados por via oral, sob a forma de comprimidos, cápsulas ou xaropes [FPP, 2017].

Quando não tratada adequadamente a **TB** mostra-se fatal [CDC, 2017].

A expressão da TB ao nível global em 2015

Segundo o *Global Tuberculosis Report 2016* [WHO, 2016] da Organização Mundial da Saúde, em 2015 surgiram cerca de 10.4 milhões de novos casos de **TB** em todo o mundo. Destes, 5.9 milhões ($\sim$ 56%) correspondiam a homens, 3.5 milhões ($\sim$ 34%) a mulheres e 1.0 milhão ($\sim$ 10%) a crianças. No mesmo relatório a OMS indica que cerca de 11% dos novos casos correspondem a pessoas infetadas com o HIV e que apenas seis países representam 60% do total de novos casos, sendo eles: a Índia, a Indonésia, a China, a Nigéria, o Paquistão e a África do Sul.

Estima-se que no ano de 2015 tenham morrido cerca de 1.8 milhões de pessoas um pouco por todo o mundo devido à TB.

A Organização Mundial da Saúde diz-nos ainda que entre os anos de 2000 e 2015 o número de mortes por tuberculose diminuiu cerca de 22%, lembrando ainda que esta continua a ser uma das 10 principais causas de morte em todo o mundo.

A expressão da TB em Portugal em 2014

Tendo em conta o último reporte sobre a TB em Portugal apresentado pela Direção Geral da Saúde (DGS), referente ao ano de 2014, temos que nesse ano apareceram 2080 novos casos de **TB**, o que corresponde, aproximadamente, a 20/100 000 habitantes.

Segundo o mesmo reporte, do ano de 2013 para o ano de 2014 verificou-se uma diminuição de 5% no número de novos casos. Foi ainda verificado que mais de 74% dos doentes com TB foram também diagnosticados com HIV [DGS, 2015].

Durante o ano de 2014 morreram 105 pessoas durante o tratamento, sendo que 955 dos doentes completaram o tratamento, dos quais 82.5% com sucesso [DGS, 2015].

7.3.2 Apresentação dos dados

O conjunto de dados que foi utilizado na aplicação que agora descrevemos, corresponde à leitura de 26 variáveis em indivíduos aos quais foi diagnosticada TB em Portugal. Os níveis de expressão destas 26 variáveis foram medidos em 20591 indivíduos.

Na tabela que se apresenta de seguida, é possível ver o nome de cada uma das variáveis que compõem os dados e uma curta descrição de cada uma.

Variável	**Descrição**
Sexo	Sexo do indivíduo (Masculino, Feminino).
Idade	Idade do indivíduo (anos).
País de origem	País de origem do indivíduo (Portugal, Angola, ...).
Histórico de Reclusão	O indivíduo tem histórico de reclusão (0: Não, 1: Sim).
Rastreio	Deteção da doença (Rastreio passivo, outro rastreio, ...).
Tempo para a primeira consulta	Tempo decorrido entre a deteção da doença e a primeira consulta (dias).
Insuficiência Renal Crónica em Diálise	O indivíduo apresenta insuficiência renal crónica em diálise (Sim, Não).
Linfomas ou Doenças Mieloproliferativas	O individuo apresenta Linfomas ou Doenças Mieloproliferativas (Sim, Não).
Infeção por VIH	Portador de HIV-SIDA (Sim, Não).
Doença Inflamatória Articular	Apresenta doença inflamatória articular (Sim, Não).
Outra Doença do Interstício	Apresenta outra doença do interstício (Sim, Não).
Diabetes	O individuo apresenta diabetes (Sim, Não).
Silicose	Apresenta silicose (Sim, Não).
Doença hepática	O indivíduo apresenta doença hepática (Sim, Não).
Neoplasia do pulmão	Apresenta Neoplasia pulmonar (Sim, Não).
Sarcoidose	Apresenta Sarcoidose (Sim, Não).
Neoplasia de outros órgãos	Apresenta Neoplasia de outros órgão (Sim, Não).
DPOC	Indivíduo apresenta Doença Pulmonar Obstrutiva Crónica (Sim, Não).
Dependência de álcool	Indivíduo dependente de bebidas alcoólicas (Sim, Não).
Dependência de drogas injetáveis	Indivíduo apresenta dependência de drogas injetáveis (Sim, Não).
Principal localização da doença	Principal localização da doença (Pulmonar, Pleural, ...).
RX do Torax	Resultado do Ráio-X Torácico (Normal, Cavitada, Não cavitada, ...).
Número de tratamentos anteriores	Número de tratamentos à tuberculose já efetuados pelo indivíduo (0, >0).

HR	O indivíduo apresenta multirresistência (Sim, Não).
Outcome	Resultado do período de tratamento (0: Bom outcome, 1: Mau outcome, ET-TF: Transferência ou em tratamento).
Ano	Ano em que foi diagnosticada a doença ao indivíduo (2008, 2009, ..., 2015).

Tabela 7.1: Descrição das variáveis do conjunto de dados em estudo na aplicação.

7.3.3 O problema correspondente aos dados

Como já foi referido na tabela 7.1, na matriz dos dados a variável **Outcome** representa o resultado do tratamento à Tuberculose. Esta variável é categórica, estando dividida em três categorias:

i. **0** - bom outcome, isto é, *o portador de TB terminou o tratamento*;

ii. **1** - mau outcome, isto é, *o portador de TB faleceu durante o período de tratamento, pelo que, não o terminou*; e

iii. **ET-TF** - *o portador de TB ainda se encontra em tratamento e/ou foi transferido.*

As restantes variáveis guardam informação sobre caraterísticas, histórico e sintomas dos indivíduos que padecem ou padeceram da doença.

Deste modo, tendo em conta os nossos objetivos, o nosso problema passa por modelar se o portador da doença sobrevive ou não ao tratamento, utilizando para tal a informação sobre as caraterísticas, o histórico e os sintomas dos indivíduos.

7.3.4 Pré-processamento dos dados

Tendo em conta que os nossos objetivos passam por aplicar os vários modelos de regressão logística estudados, e como para tal a variável resposta apenas pode ter duas categorias, uma representado o sucesso e a outra o insucesso do acontecimento em estudo, então apenas consideramos para os efeitos da aplicação as observações para as quais a variável **Outcome** apresenta as categorias **0** (*bom outcome*) e **1** (*mau outcome*), ou seja, descartamos do estudo as observações correspondentes a indivíduos ainda em tratamento ou que tenham sido transferidos [2]. Note-se que para os indivíduos que apresentam **Outcome ET-TF** desconhecemos qual o desfecho do tratamento, pelo que não faz sentido estes serem incluídos na análise.

Assim, ao conjunto de dados foram retiradas as observações com **Outcome ET-TF**, sendo o conjunto de dados resultante, e com o qual passaremos a trabalhar de seguida, composto pelas mesmas 26 variáveis, mas contendo apenas 19519 observações.

[2]Os indivíduos **transferidos** correspondem a indivíduos que estavam reclusos e que ou foram transferidos de estabelecimento prisional ou foram colocados em liberdade, tendo-se perdido o seu acompanhamento.

7.4 Descrição dos dados

Ao nível da análise dos dados, começamos por fazer a sua descrição utilizando a mediana, o mínimo e o máximo como medidas descritivas para as variáveis contínuas [3] e as frequências absoluta e relativa para as variáveis categóricas. A descrição resultante da aplicação destas medidas pode ser vista na tabela do **Anexo A**.

Ao lermos os resultados de tal descrição, notamos que algumas das variáveis necessitam de ser recategorizadas e que existe um conjunto de variáveis binárias com elevados problemas de separabilidade.

As variáveis às quais identificamos problemas de separabilidade são: **Insuficiência Renal Crónica em Diálise**, **Linfomas ou Doenças Mieloproliferativas**, **Inflamatória articular**, **Outra doença do intersticio**, **Diabetes**, **Silicose**, **Doença hepática**, **Neoplasia do pulmão**, **Sarcoidose**, **Neoplasia de outros orgãos** e **DPOC**.

A generalidade das variáveis enumeradas correspondem a doenças que podem aparecer simultaneamente que a **Tuberculose**, ou seja, a **Comorbilidades**.

Deste modo, decidimos criar uma única variável que agrupe a informação destas variáveis, designando-a por **Comorbilidades**. Ao criarmos esta variável pretendemos diminuir o risco de observar separabilidade nos dados aos quais iremos aplicar os modelos de regressão.

Nesta nova variável, cada observação, está categorizada com **0** ou com **1**. A categoria **1** representa as observações cujos respetivos indivíduos apresentam pelo menos uma das doenças, isto é, pelo menos uma comorbilidade. Já a categoria **0** reúne todas as observações cujos indivíduos não apresentam nenhuma das doenças.

Na tabela 7.2, apresenta-se uma breve análise descritiva das várias variáveis que serão utilizados na aplicação, considerando já as variáveis recategorizadas e a variável **Comorbilidades**.

Variável	Total	Outcome	
		Bom outcome	Mau outcome
Sexo	n (%)	n (%)	n (%)
Masculino	12781 (65.480)	11240 (64.450)	1541 (74.122)
Feminino	6738 (34.520)	6200 (35.550)	538 (25.878)
Idade	n (%)	n (%)	n (%)
< 20	1007 (5.164)	977 (5.607)	30 (1.445)
$20 - 40$	6693 (34.321)	6190 (35.524)	503 (24.229)
$40 - 60$	6991 (35.849)	6349 (36.436)	642 (30.925)
> 60	4010 (24.665)	3909 (22.433)	901 (43.401)
País de origem	n (%)	n (%)	n (%)
Portugal	16509 (84.579)	14758 (84.622)	1751 (84.223)
Outros	3010 (15.421)	2682 (15.378)	328 (15.777)

[3] Uma vez que estas apresentam assimetrias.

Histórico de Reclusão	n (%)	n (%)	n (%)
0 (Não)	19196 (98.345)	17147 (98.320)	2049 (98.557)
1 (Sim)	323 (1.655)	293 (1.680)	30 (1.443)
Rastreio	n (%)	n (%)	n (%)
Rastreio passivo	17639 (90.368)	15728 (90.183)	1911 (91.919)
Outro rastreio	1148 (5.881)	1080 (6.193)	68 (3.271)
Desconhecido	732 (3.750)	632 (3.624)	100 (4.810)
Tempo para a primeira consulta	Md (min - max)	Md (min - max)	Md (min - max)
	34 (0 - 7536)	34 (0 - 4527)	31 (0 - 7536)
Comorbilidades	n (%)	n (%)	n (%)
0 (Não)	16530 (84.687)	15028 (86.170)	1502 (72.246)
1 (Sim)	2989 (15.313)	2412 (13.830)	494 (27.754)
Infeção por VIH	n (%)	n (%)	n (%)
Não	17374 (89.011)	15751 (90.315)	1623 (78.066)
Sim	2145 (10.989)	1689 (9.685)	456 (21.934)
Depend. de Álcool	n (%)	n (%)	n (%)
Não	16082 (82.392)	14624 (83.853)	1458 (70.130)
Sim	2386 (12.224)	2008 (11.514)	378 (18.182)
Desconhecido	1051 (5.384)	808 (4.633)	243 (11.688)
Dependência de drogas injetáveis	n (%)	n (%)	n (%)
Não	17313 (88.698)	15727 (90.170)	1586 (76.287)
Sim	1202 (6.158)	939 (5.400)	263 (12.650)
Desconhecido	1004 (5.144)	774 (4.430)	230 (11.063)
Principal localização da doença	n (%)	n (%)	n (%)
Pulmonar	14082 (72.145)	12562 (72.030)	1520 (73.112)
Pleural	1640 (8.402)	1479 (8.481)	161 (7.744)
Outra	3797 (19.453)	3399 (19.490)	398 (19.144)
Raio-X Torax	n (%)	n (%)	n (%)
Cavitada	6988 (35.800)	6385 (36.611)	603 (29.004)
Não Cavitada	8065 (41.319)	7115 (40.797)	950 (45.695)
Normal	2934 (15.032)	2696 (15.459)	238 (11.448)
Desconhecida	1532 (7.849)	1244 (7.133)	288 (13.853)
N. tratamentos anteriores	n (%)	n (%)	n (%)
0	17822 (91.306)	15991 (91.692)	1831 (88.071)
> 0	1697 (8.694)	1449 (8.308)	248 (11.929)
ANO	n (%)	n (%)	n (%)

< 2014	15505 (79.435)	13920 (79.817)	1585 (76.239)
2014+	4014 (20.565)	3520 (20.183)	494 (23.761)
HR	n (%)	n (%)	n (%)
0	19345 (99.109)	17307 (99.237)	2038 (98.028)
1	174 (0.891)	133 (0.763)	41 (1.972)

Tabela 7.2: Breve análise descritiva das variáveis do conjunto de dados utilizados na aplicação.

É importante referir que das 16 [4] variáveis que serão utilizadas na aplicação aos modelos de regressão, uma é continua, sendo as restantes 15 categóricas.

7.5 Modelos de regressão

O que apresentamos de seguida corresponde à aplicação dos vários modelos de regressão logística anteriormente abordados ao conjunto de dados. Com esta aplicação pretendemos determinar, para cada modelo de regressão, o modelo que melhor se ajusta aos dados, com a finalidade de mais à frente podermos comparar os vários modelos de regressão estudados.

7.5.1 Modelo de regressão logística usual

O primeiro modelo que aplicamos os dados foi o **modelo de regressão logística usual**.

Começamos por dividir aleatoriamente o nosso conjunto de dados em dois, um com 70% e o outro com 30% da totalidade das observações. Ao conjunto que contém 70% das observações designamos por **Conjunto de Treino** e será com base neste que iremos ajustar o nosso modelo. Ao conjunto que contém 30% do total de observações designamos por **Conjunto de Teste**, sendo com base neste que iremos avaliar os modelo ajustado.

Posto isto, no **R**, pedimos que fosse obtido o **modelo de regressão logística usual** utilizando para o ajustamento todas as variáveis do Conjunto de Treino.

No **Anexo B**, é possível ver o resultado obtido com o '*summary( )*' do modelo completo.

Em tais resultados é possível verificar que, tal como já era previsível, várias das variáveis não se mostram significativas para o modelo. Assim, fomos passo a passo, removendo a variável com pior significância com o objetivo de encontrar o melhor modelo.

Ao longo deste processo fomos avaliando os vários modelos que íamos obtendo, utilizando um conjunto de medidas.

Na tabela que se segue, é possível ver os valores das medidas utilizadas para os vários modelos obtidos.

[4] Uma variável resposta e 15 variáveis explicativas.

Modelo	Valor - p	AIC	Acurácia	Kappa	AUC
glm1	< 0.001	5944.381	0.906	0.040	0.511
glm2	< 0.001	5942.416	0.906	0.040	0.511
glm3	< 0.001	5941.977	0.906	0.040	0.511
glm4	< 0.001	5939.079	0.907	0.044	0.513
glm5	< 0.001	**5939.693**	**0.907**	**0.045**	**0.513**
glm6	< 0.001	5959.903	0.907	0.040	0.512
glm7	< 0.001	5961.343	0.907	0.040	0.512

Tabela 7.3: Valores de algumas medidas dos modelos avaliados com a regressão logística usual.

Veja-se pela tabela anterior que as medidas utilizadas foram o **Valor-p**, o **AIC**, a **Acurácia**, o **Kappa** e o **AUC**. É de relevo salientar que as três últimas medidas enumeradas são baseadas na informação da matriz de confusão resultante da aplicação do modelo ao conjunto de teste.

Para os sete modelos avaliados, o **Valor-p** permaneceu < 0.001, não sendo, claro está, uma medida que nos permita diferenciar os vários modelos. É importante aqui referir que este valor resulta da comparação do modelo em estudo com o modelo nulo, permitindo-nos recusar a hipótese da igualdade entre estes modelos.

Por seu turno, o critério **AIC**, isto é, *Akaike Information Criteria*, foi variando de modelo para modelo, sendo como tal, uma medida com alguma materialidade para avaliar qual destes é o melhor.

Segundo a literatura, o critério **AIC**, foi proposto pelo estatístico japonês Hirotugu Akaike em 1973, correspondendo à seguinte expressão:

$$AIC = 2K - 2log(L(\widehat{\theta}|y)), \tag{7.1}$$

onde, K corresponde ao número de parâmetros estimáveis e $log(L(\widehat{\theta}|y))$ ao valor da função de log-verosimilhança avaliada com os valores estimados para o modelo. Comparando dois modelos, o melhor, segundo este critério, será o que apresenta o menor valor [Snipes & Taylor, 2014].

Assim, tendo em conta os valores apresentados na tabela 7.3, o melhor modelo segundo este critério é o *glm4*. Contudo, note-se que o valor obtido para o modelo *glm5* é muito semelhante ao obtido para o modelo *glm4*.

Outra das medidas utilizadas foi a **Acurácia**. Esta medida tem por objetivo avaliar a exatidão e a precisão do modelo, sendo obtido a partir da informação da matriz de confusão resultante da aplicação do conjunto de teste.

Comparando dois modelos, considera-se, tendo em conta esta medida, que o melhor modelo é aquele que apresenta o valor mais elevado.

Acontece que os sete modelos apresentam valores de acurácia semelhantes, diferenciando-se apenas ao nível das centésimas. O valor mais elevado de entre os obtidos é 0.907, para os

modelos *glm4*, *glm5*, *glm6* e *glm7*.

O valor de **Kappa** é uma medida de concordância que mede o grau de concordância, entre os valores previstos e os observados, além do que seria esperado pelo acaso. Esta medida tem como valor máximo 1, representando este a concordância total. Por seu turno, valores próximos ou até menores que 0 indicam a falta de concordância [Baltar & Okano, 2017]. Tal como a **Acurácia**, também o valor de **Kappa**, é baseado na informação da matriz de confusão.

Note-se que o valor de **Kappa** é muito reduzido para os vários modelos avaliados. Ainda assim, o modelo *glm5* é o que apresenta o melhor resultado.

Por fim, a última medida considerada foi o valor da **AUC**. **AUC** é a sigla da expressão inglesa **A**rea **U**nder the **C**urve. Esta é uma medida que se baseia na área da **Curva ROC**. As **Curvas ROC** são construidas a partir da informação das matriz de confusão, servindo para avaliar os modelos de regressão que deram origem às ditas matrizes de confusão [Lobo & et al., 2008].

É importante referir que um valor de **AUC** próximo de 1 corresponde a um bom modelo de regressão, uma vez que corresponde a uma concordância elevada entre os valores previstos e os valores observados. Por outro lado, valores próximos de 0.5 correspondem a uma concordância reduzida.

Note-se que na tabela 7.3, os valores de **AUC** são para os vários modelos muito reduzidos, o que nos permite dizer que a concordância entre os resultados estimados pelos modelos e os valores reais é muito reduzida. Ainda assim, quando observamos o valor às milésimas, podemos concluir que, para esta medida, os modelos *glm4* e *glm5* são os que apresentam os melhores resultados.

Posto isto, e tendo em conta os valores obtidos para as várias medidas utilizadas, podemos concluir que o melhor modelo obtido com a aplicação dos dados ao **modelo de regressão logística usual** é o *glm5*.

Deste modo, pedimos ao **R** que nos representasse graficamente os resíduos referentes ao modelo *glm5*. Apresentamos tais representações gráficas na figura 7.1.

Pelas representações gráficas da figura 7.1, podemos ver que não existem resíduos estandardizados considerados muito grandes, isto é, '> 3.3'. Também é possível ver que os valores dos resíduos estandardizados são na sua maiorias inferiores a zero.

Deixa-se de seguida a matriz de confusão resultante da aplicação do conjunto de teste ao modelo *glm5*.

```
          Reference
Prediction    0    1
         0 3877  392
         1    7   11
```

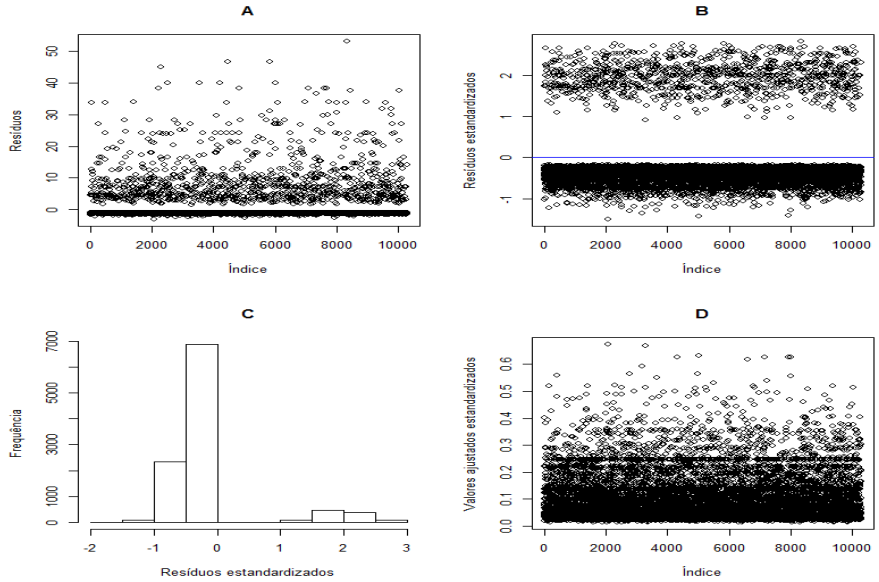


Figura 7.1: Representação gráfica dos resíduos (A), dos resíduos estandardizados (B), do histograma dos resíduos estandardizados (C) e dos Valores ajustados standardizados (D).

Finda-se esta parte dos resultados com o seguinte recorte, no qual podem ser vistas as variáveis que fazem parte do modelo *glm5*, bem como os respetivos coeficientes.

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -3.30588    0.29444 -11.227  < 2e-16
I(Sexo)1                         0.35591    0.08422   4.226 2.38e-05
I(Idade)>60                      1.63867    0.26415   6.204 5.52e-10
I(Idade)20 - 40                  0.33268    0.26717   1.245  0.21305
I(Idade)40 - 60                  0.45606    0.26598   1.715  0.08641
I(Pais.de.Origem)Portugal       -0.43138    0.09704  -4.445 8.77e-06
I(Dep..de.alcool)1               0.52796    0.09769   5.405 6.49e-08
I(Dep..de.alcool)Desconhecido    0.68588    0.17176   3.993 6.52e-05
I(Hx.reclusao)1                 -0.95695    0.36667  -2.610  0.00906
I(Infeccao.por.VIH)1             0.91503    0.10698   8.553  < 2e-16
I(Dep..drogas.ev)1               0.93285    0.13486   6.917 4.60e-12
I(Dep..drogas.ev)Desconhecido    0.48471    0.18822   2.575  0.01002
I(RX.Torax)Desconhecida          0.59689    0.13307   4.486 7.27e-06
I(RX.Torax)Nao Cavitada          0.13080    0.08058   1.623  0.10454
I(RX.Torax)Normal               -0.16646    0.12518  -1.330  0.18362
I(N..tratamentos.anteriores)0   -0.21670    0.10927  -1.983  0.04734
I(HR)1                           1.09242    0.25743   4.244 2.20e-05
I(Comorbilidades)1               0.70910    0.08527   8.316  < 2e-16
```

É de extremo relevo deixar uma nota de que se fossem aqui apresentados os efeitos brutos de cada uma das variáveis numa regressão simples, estes seriam para todas elas significativos, dado o elevado tamanho da amostra.

7.5.2 Modelo de regressão logística condicional em estudos de caso-controlo

Como foi dito aquando da descrição teórica, para a aplicação do modelo de regressão logística condicional em estudos de caso-controlo é essencial que cada **caso** seja correspondido com pelo menos um **controlo**.

Para a construção dos conjuntos correspondentes é essencial conhecermos as melhores variáveis a utilizar como variáveis de confundimento. Tendo em conta a opinião dos profissionais da área, decidiu-se utilizar como variáveis de confundimento o **Sexo**, a **Idade** [5], a **Dependência de Álcool**, a **Dependência de drogas injetáveis** e a **Infeção por VIH**.

Posto isto, foi necessário achar uma função que, ao nível do **R**, fizesse o emparelhamento das observações. Acontece que da pesquisa efetuada, as funções encontradas eram bastante especificas ao conjunto de dados ou à forma destes. Assim, decidiu-se criar, por conta própria, um *script* com o conjunto de instruções a efetuar o emparelhamento.

Tal conjunto de instruções pode ser visto no **Anexo C**.

Salienta-se aqui que a bibliografia não nos indica o número certo de controlos a emparelhar com cada caso, dizendo-nos apenas que este número, M, deve variar entre um e cinco.

Assim, utilizando o conjunto de instruções apresentadas no **Anexo C**, pedimos ao software que nos fizesse todos os emparelhamentos, para os vários valores de **M**. Como resultado ficamos com cinco matrizes de dados, cada uma delas correspondendo aos emparelhamentos de uma caso com M controlos, $M \in \{1, 2, 3, 4, 5\}$.

O passo dado com o emparelhamento dos casos com **M** controlos constitui, muito possivelmente, o passo mais difícil e importante nesta aplicação.

De seguida, vamos aplicar o modelo de regressão às várias matrizes de emparelhamento obtidas. No caso do modelo de regressão logística condicional em estudos de caso-controlo pretendemos determinar o melhor modelo que representa os dados, ao qual está implicitamente associada a determinação do melhor número de controlos a emparelhar com cada caso.

Assim, tal como foi feito para o **modelo de regressão logística usual**, também aqui dividimos o conjunto de dados referentes a cada emparelhamento em dois. Um conjunto de treino e um conjunto de teste, nas proporções de 70% e 30%, respetivamente.

Começamos, para cada valor de M, por pedir o modelo completo. De seguida, foi-se, passo a passo, retirando a variável menos significativa do modelo, anotando-se os valores do conjunto de medidas já comentadas para cada um dos modelos.

[5] Motivo pelo qual festa variável foi categorizada.

Na tabela 7.4 é possível ver os valores correspondentes ao melhor modelo para cada valor de M. Os resultados obtidos nos vários modelos, para cada valor de M, podem ser vistos no **Anexo D**.

M	Modelo	Valor-p	AIC	Acurácia	Kappa	AUC
1	*clogit14*	< 0.001	977.512	0.521	0.063	0.532
2	*clogit26*	< 0.001	1623.814	0.601	0.070	0.535
3	*clogit34*	< 0.001	2069.650	0.443	0.049	0.540
4	*clogit46*	< 0.001	2408.398	0.627	0.063	0.540
5	***clogit58***	< 0.001	**2510.472**	**0.661**	**0.068**	**0.546**

Tabela 7.4: Valores de algumas medidas avaliadas no melhor modelo obtido com a regressão logística condicional em estudos de caso controlo, para cada emparelhamento 1:M.

Olhando para a tabela anterior, podemos evidenciar que ao nível que o valor de M aumenta, também o valor de **AIC** aumenta. Tal justifica-se facilmente com o facto de o valor da log-verosimilhança estar dependente do tamanho da amostra, e esta do valor de M.

Em segundo lugar, note-se que é com o valor de $M = 3$ que é tido o pior valor de **acurácia**, sendo, para esta medida, o emparelhamento 1:5 o que apresenta o melhor resultado.

Ao nível do valor de **Kappa**, o modelo que apresenta o valor mais elevado é o melhor modelo obtido para o emparelhamento 1:2.

Quando olhamos para os resultados ao nível da variável **AUC**, podemos ver que o modelo *clogit58* é o que apresenta o melhor valor.

Nas representações da figura 7.2, é possível ver-se alguns dos resultados anteriores em *plot*.

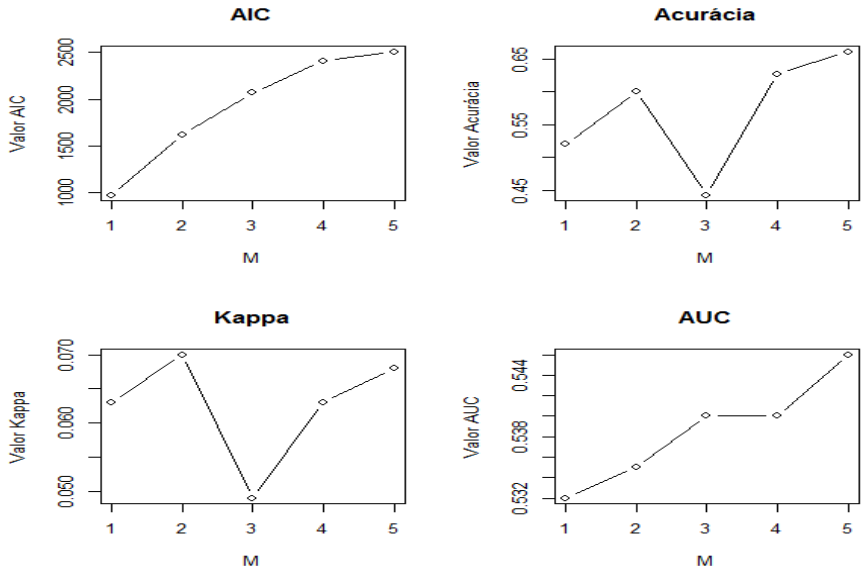


Figura 7.2: Representação gráfica dos valores de **AIC**, **Acurácia**, **Kappa** e **AUC** obtidos para os melhores modelos de cada emparelhamento 1:M.

Note-se, pelas representações, que o melhor modelo obtido para o emparelhamento 1:5 é o que tem os melhores resultados para a **Acurácia** e para a **AUC**, tendo o segundo melhor valor para a medida **Kappa**. Para esta última, o modelo que apresenta o valor mais elevado é o melhor modelo resultante do emparelhamento de cada **caso** com dois **controlos**.

Assim, podemos concluir que o modelo que melhor se ajusta aos nossos dados é o *clogit58* - o melhor de entre os modelos que resultaram do emparelhamento de cada **caso** com cinco **controlos**.

De seguida, na figura 7.3, apresentam-se representações gráficas dos resíduos do modelo **clogit58**.

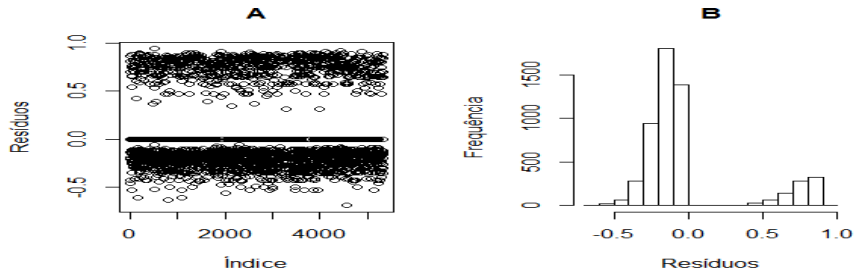


Figura 7.3: Representação gráfica dos resíduos (A) e do histograma dos resíduos (B) para o modelo *clogit58*.

Pode-se notar, nas representações gráficas da figura 7.3, que os resíduos são em módulo inferiores a 1, e que se concentram essencialmente abaixo de zero.

De seguida apresentamos a matriz de confusão referente ao modelo *clogit58* [6]:

```
          Reference
Prediction    0    1
         0 1394  234
         1  551  140
```

Note-se que utilizado o modelo *clogit58* foi possível acertar no **Outcome** de 140 das observações, o que corresponde a mais de 37% dos **casos**. Como senão, temos os 551 controlos (cerca de 28% do total de controlos no conjunto de teste) para os quais foi errada a previsão do **Outcome**.

O modelo *clogit58* é bastante simples, contemplando apenas três variáveis. De seguida, apresentam-se os resultados referentes aos coeficientes desse modelo.

```
                             coef exp(coef) se(coef)      z Pr(>|z|)
I(Pais.de.Origem)Portugal -0.42879   0.65129  0.11938 -3.592 0.000328
I(HR)                      0.92065   2.51093  0.35900  2.564 0.010333
I(Comorbilidades)          0.80026   2.22613  0.09229  8.671  < 2e-16
```

Lembro, tal como foi concluído na exposição teórica deste método que o modelo não contempla um coeficiente livre.

[6] Não se verifica a proporção 1 : 5 devido à existência de *missing values* dos preditores.

7.5.3 Modelo de regressão logística com correção do viés, usando correção apriori e pesos

De modo a facilitar a nomenclatura do modelo de regressão logística com correção do viés, usando correção apriori e pesos, proposto por King e Zeng, chamemos-lhe *Relogit*.

Tal como foi indicado na explicação teórica deste modelo de regressão, quando as dimensões dos dados são consideráveis a melhor correção a utilizar é a ponderada.

Não podemos esquecer que este modelo de regressão baseia-se em dados com emparelhamento caso-controlo. Desta feita, e como não há uma condição que nos indique quantos controlos devem ser emparelhados com cada caso, decidiu-se utilizar como amostra o conjunto de dados resultante do emparelhamento 1:5.

Para além disto é essencial considerar um valor para τ, isto é, a proporção de eventos de interesse que ocorrerem na população. Assim, como o nosso conjunto de dados corresponde aos registo de todos os diagnósticos de TB em Portugal, considerou-se que a proporção de mortes por TB em Portugal era de 2079/19519, isto é, a proporção de mortes no conjunto de dados.

De um modo idêntico aos restantes modelos de regressão estudados, começou-se por dividir os dados em dois conjuntos, um de treino, com 70% das observações, e um de teste, com as restantes 30%.

Posto isto, pediu-se ao *software* que nos calcula-se o modelo completo. De seguida foi-se, passo a passo, retirando a variável menos significativa do modelo, anotando-se os valores do conjunto de medidas já utilizadas nos modelos de regressão anteriores.

Os resultados são os que se apresentam de seguida:

Modelo	**Valor - p**	**AIC**	**Acurácia**	**Kappa**	**AUC**
relogit1	< 0.001	4724.884	0.839	0	0.5
relogit2	< 0.001	4722.891	0.839	0	0.5
relogit3	< 0.001	4718.985	0.839	0	0.5
relogit4	< 0.001	**4718.013**	**0.839**	**0**	**0.5**
relogit5	< 0.001	4716.083	0.839	0	0.5
relogit6	< 0.001	4722.022	0.839	0	0.5
relogit7	< 0.001	4721.822	0.839	0	0.5

Tabela 7.5: Valores de algumas medidas dos modelos avaliados para o modelo Relogit.

Note-se, na tabela anterior, que os valores para as várias medidas são iguais para a generalidade dos modelos, diferenciando-se apenas ao nível do valor de **AIC**.

É de destacar o facto de todos os modelos apresentarem um valor nulo na variável **Kappa** e de apresentarem o valor de 0.5 para a variável **AUC**.

Voltando à variável **AIC**, temos que o valor mais baixo foi o obtido para o modelo *relogit4*, pelo que o consideramos como sendo o nosso melhor modelo.

Na figura 7.4 apresentam-se as representações gráficas dos resíduos do modelo **relogit4**.

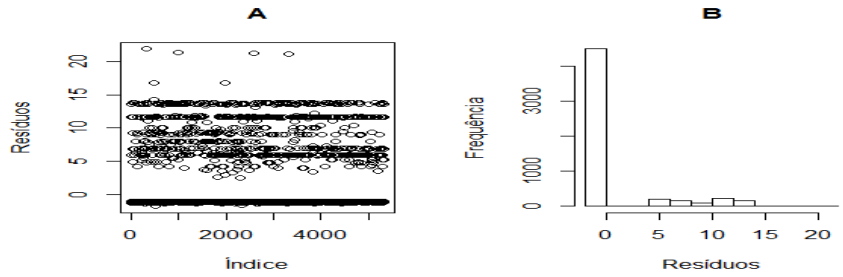


Figura 7.4: Representação gráfica dos resíduos (A) e do histograma dos resíduos (B) para o modelo *relogit4*.

Em tais representações é fácil de observar que existem valores de resíduos bastante elevados. Contudo, a generalidade dos resíduos têm um valor muito próximo de zero.

De seguida apresenta-se a matriz de confusão obtida para o modelo *relogit4*:

```
          Reference
Prediction    0    1
         0 1945  374
         1    0    0
```

Note-se pela matriz anterior que o modelo traduz a previsão de todas as observações na categoria **0**, errando, portanto, todas as observações que contemplavam a categoria **1**.

De seguida apresentam-se as variáveis que fazem parte do modelo *relogit4* e os respetivos coeficientes:

```
Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -1.9558947  0.2158521  -9.061  < 2e-16
I(Pais.de.Origem)Portugal      -0.4281681  0.1256144  -3.409 0.000653
I(Rastreio)Desconhecido         0.2618400  0.2342216   1.118 0.263603
I(Rastreio)Outro_rastreio      -0.5643143  0.4224885  -1.336 0.181650
Tempo.para.a.primeira.consulta  0.0001528  0.0002562   0.597 0.550831
I(RX.Torax)Cavitada             0.0131137  0.1501106   0.087 0.930385
I(RX.Torax)Desconhecida         0.4687111  0.1988657   2.357 0.018427
I(RX.Torax)Nao Cavitada         0.1857737  0.1409906   1.318 0.187627
I(N..tratamentos.anteriores)0  -0.1689352  0.1441907  -1.172 0.241354
I(HR)                           0.9894839  0.3703985   2.671 0.007553
I(Comorbilidades)               0.7634063  0.0963710   7.922 2.35e-15
```

7.5.4 Modelo de regressão logística de Firth

Começou-se por dividir o conjunto de dados em dois conjunto distintos, um conjunto de Treino e um conjunto de Teste, utilizando as mesmas proporções que para os restantes modelos de regressão abordados.

De seguida, o procedimento utilizado é idêntico ao seguido para os restantes modelos aplicados.

Deste modo, começamos por pedir ao *software* que nos ajustasse o modelo completo. Partindo desse modelo, fomos tentar achar o modelo, que seguindo o modelo de regressão de Firth, melhor se ajusta aos nossos dados.

Assim, passo a passo, foram sendo eliminadas, uma a uma, as variáveis que apresentavam menor significância.

Na tabela que se segue encontram-se resumidos os resultados obtidos para as medidas avaliadas nos vários modelos considerados para a regressão logística de Firth.

Modelo	**Valor - p**	**AIC**	**Acurácia**	**Kappa**	**AUC**
brglm1	< 0.001	5948.355	0.661	0.068	0.546
brglm2	< 0.001	5946.453	0.907	0.016	0.505
brglm3	< 0.001	5944.772	0.907	0.016	0.505
brglm4	< 0.001	**5943.585**	**0.907**	**0.021**	**0.506**
brglm5	< 0.001	5942.694	0.907	0.017	0.505
brglm6	< 0.001	5943.054	0.907	0.016	0.505
brglm7	< 0.001	5943.608	0.907	0.017	0.505

Tabela 7.6: Valores de algumas medidas dos modelos avaliados com a regressão logística de Firth.

Note-se pelos valores apresentados na tabela anterior que a generalidade dos modelos avaliados apresentam o mesmo valor de **acurácia**, no caso 0.907. Ao nível do valor **Kappa** é de destacar a diferença da medida obtida para o modelo completo, comparativamente com os restantes modelos. Também se destacam fortemente os valores do modelo completo comparativamente com os restantes modelos, na medida **AUC**. Contudo, a medida **acurácia** é muito reduzida.

Deste modo, o nosso foco voltar-se-á para o modelo *brglm4*, que não tendo os melhores valores de **Kappa** e **AUC**, ainda assim, destaca-se positivamente dos demais.

Na figura 7.5, é possível ver as representações gráficas dos resíduos do modelo *brglm4*.

Pelas representações gráficas, podemos ver que não existem resíduos estandardizados '> 3.3'. Também é possível ver que os valores de resíduos estandardizados são na sua maiorias inferiores a zero.

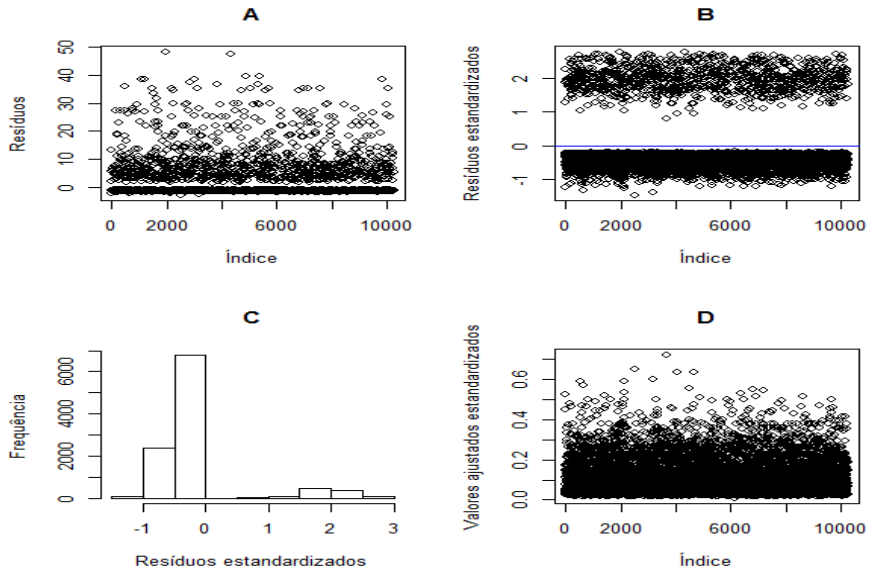


Figura 7.5: Representação gráfica dos resíduos (A), dos resíduos estandardizados (B), do histograma dos resíduos estandardizados (C) e dos Valores ajustados standardizados (D) do modelo *brglm4*.

Deixa-se de seguida a matriz de confusão resultante da aplicação do conjunto de teste ao modelo *brglm4*, com o objetivo de comprar os valores reais com os valores previstos com o modelo.

```
          Reference
Prediction    0    1
         0 3933  402
         1    3    5
```

Finda-se esta parte dos resultados com os coeficientes referentes a cada variável:

```
Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -3.47418    0.32256 -10.771  < 2e-16
I(Sexo)1                          0.26056    0.08248   3.159 0.001582
I(Idade)>60                       1.62554    0.25956   6.263 3.79e-10
I(Idade)20 - 40                   0.30500    0.26280   1.161 0.245813
I(Idade)40 - 60                   0.53711    0.26136   2.055 0.039879
I(Dep..de.alcool)1                0.44229    0.09716   4.552 5.31e-06
I(Dep..de.alcool)Desconhecido     0.53277    0.17921   2.973 0.002951
I(Hx.reclusao)1                  -0.46911    0.33107  -1.417 0.156496
I(Infeccao.por.VIH)1              0.94973    0.10737   8.845  < 2e-16
I(Dep..drogas.ev)1                0.87492    0.13504   6.479 9.23e-11
I(Dep..drogas.ev)Desconhecido     0.50804    0.19157   2.652 0.008002
I(Pais.de.Origem)Portugal        -0.34527    0.09844  -3.507 0.000453
I(RX.Torax)Cavitada               0.14598    0.16096   0.907 0.364451
I(RX.Torax)Desconhecida           0.87720    0.16224   5.407 6.41e-08
I(RX.Torax)Nao Cavitada           0.33262    0.14891   2.234 0.025498
```

```
I(N..tratamentos.anteriores)0             -0.16527    0.10983  -1.505 0.132382
I(HR)1                                     0.68937    0.30606   2.252 0.024297
I(Comorbilidades)1                         0.73177    0.08518   8.590  < 2e-16
I(TB.Doenca.Localizacao.Principal)Outra   -0.18065    0.12724  -1.420 0.155670
I(TB.Doenca.Localizacao.Principal)Pleural -0.16226    0.13207  -1.229 0.219228
```

7.6 Comparação dos vários modelos

De modo a facilitar a nomenclatura dos vários modelos, daqui por diante, designaremos por **GLM** o melhor modelo obtido com a regressão logística usual, por **CLOGIT** o melhor modelo obtido com a regressão logística condicionada em estudos de caso-controlo, por **RELOGIT** o melhor modelo obtido utilizando a abordagem proposta por King e Zeng e por **BRGLM** o melhor modelo obtido para a regressão logística de Firth.

Na representação gráfica que se segue encontram-se as representações das **Curvas ROC** para os vários modelos de regressão em comparação, considerando para tal o melhor modelo obtido para cada um.

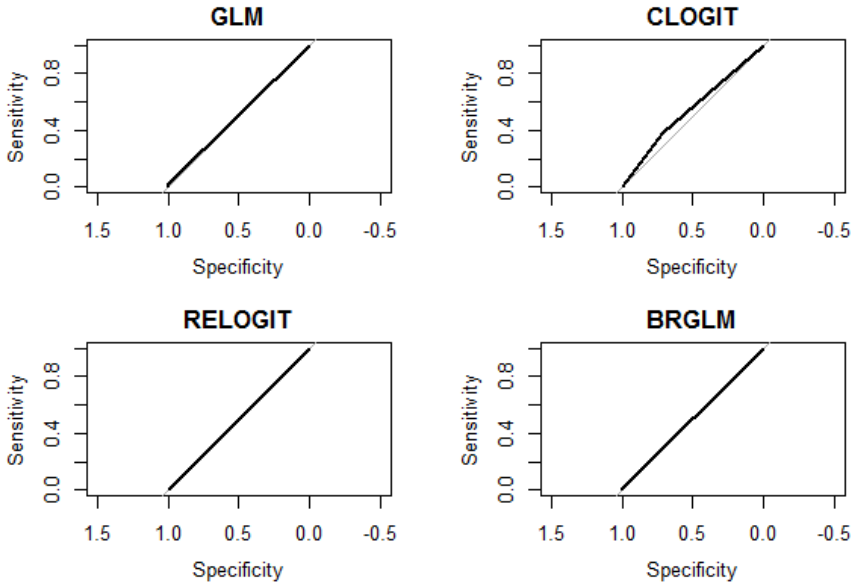


Figura 7.6: Representação gráfica das **Curvas ROC** para os vários modelos em comparação.

Note-se que em nenhuma das representações a curvatura é a desejável, ou seja, nenhum dos modelos foi capaz de aprender os dados como era pretendido. As representações obtidas já eram de esperar, tendo em conta os valores que foram sendo obtidos para a medida **AUC**.

Tal como os gráficos sugerem, o melhor valor de **AUC** foi obtido para o modelo **CLOGIT**. Tal é-nos sugerido, uma vez que é a única representação na qual podemos observar uma pequena curvatura. Nos restantes a curva é praticamente um segmento de reta, pelo que o valor de **AUC** está muito próximo de 0.5.

Tais conclusões, retiradas a partir das representações gráficas da figura 7.6, são confirmadas pelos gráficos apresentados na figura abaixo.

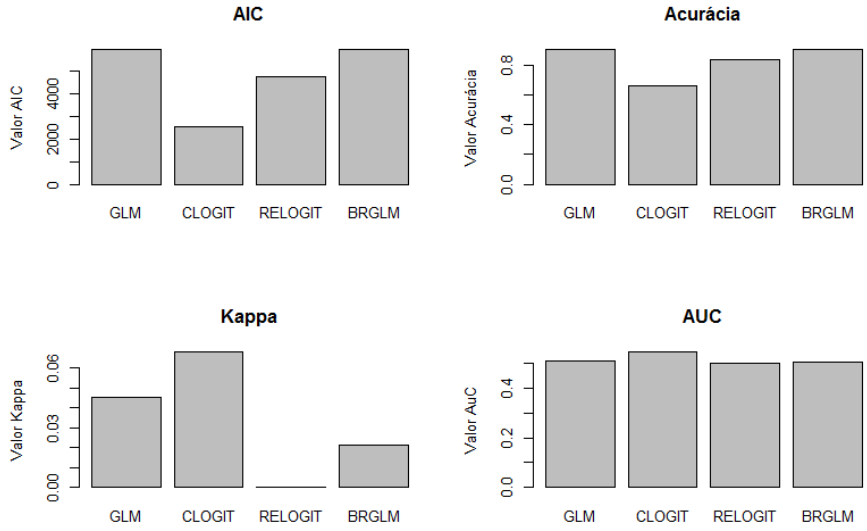


Figura 7.7: Gráficos dos valores de várias medidas, para os vários modelos em comparação.

Quando olhamos para as representações da figura anterior, o primeiro resultado que nos salta à vista é o valor nulo na medida **Kappa** para o modelo **RELOGIT**.

Acontece que se voltarmos ao capítulo anterior, podemos observar na matriz de confusão referente ao modelo **RELOGIT** que este previu todos os casos como sendo controlos. É este o motivo pelo qual o valor de **Kappa** é nulo.

Assim, desde já, o modelo **RELOGIT** fica descartado de ser o melhor modelo de regressão logística para dados com eventos raros, de entre os que foram abordados.

No que se refere ao modelo de regressão logística de Firth (**BRGLM**) este é, a par do modelo **GLM**, o que apresenta o maior valor de **Acurácia**. Contudo, quando olhamos para a sua **Curva ROC** percebemos que o valor de **AUC** está muito próximo de 0.5, isto é, do limite inferior do intervalo de valores possíveis para esta medida.

Se olharmos para a matriz de confusão do modelo **BRGLM**, percebemos que este apenas previu corretamente 5 dos 407 casos do conjunto de teste nele utilizado, tendo errado na previsão de 402 casos e na previsão de 3 controlos, que erradamente classificou como casos.

Sendo melhor que o modelo **RELOGIT**, ainda assim, o modelo **BRGLM** não é o melhor modelo para os nossos dados.

O modelo **GLM** foi um dos que apresentou melhores resultados. Este modelo, apresenta, a par do modelo **BRGLM**, o melhor valor de **Acurácia** e apresenta o segundo melhor valor de **Kappa**. Quando olhamos para a sua **Curva ROC** percebemos que o seu valor de **AUC** é muito próximo do mínimo, isto é, de 0.5.

Quando nos socorremos da sua matriz de confusão, podemos observar que o modelo acertou em 11 casos de 403, tendo errado num total de 399 observações - 7 controlos e 392 casos.

Tendo em conta o resultado para a medida **Kappa**, obtido graças à correta previsão dos 11 casos, podemos afirmar que o modelo **GLM** ajustou-se melhor aos dados que os modelos **RELOGIT** e **BRGLM**.

A representação gráfica dos valores de **AUC** apresentada na figura 7.7 permite-nos confirmar o que já tínhamos afirmado sobre o valor desta medida para o modelo **CLOGIT**, isto é, que o maior valor de **AUC** foi o obtido para este modelo.

Contudo, o modelo **CLOGIT** é, ao mesmo tempo, o modelo que apresenta o pior valor de **Acurácia**. Tal deve-se ao facto deste modelo errar na previsão de muitos "insucessos", para poder acertar nos "sucessos".

Pela respetiva matriz de confusão podemos ver que o modelo para acertar em 140 dos 374 casos que compunham o conjunto de teste, errou na previsão de 551 dos 1945 controlos.

É essencialmente devido a este erro de previsão dos controlos que o valor de **Acurácia** é com alguma expressão inferior aos dos demais modelos.

Por outro lado, quando olhamos para o valor de **Kappa** observamos que o **CLOGIT** é o modelo que apresenta o melhor resultado. Este facto deve-se essencialmente ao modelo conseguir acertar em 140 dos 374 casos.

O facto de o modelo ter conseguido acertar em mais de 37 % das observações que correspondiam a casos, constitui um resultado importante para dizer que o modelo **CLOGIT** foi, de entre os modelos avaliado, o modelo para o qual foi obtido o melhor resultado.

Capítulo 8

Conclusão

O propósito desta dissertação centrou-se essencialmente no estudo da regressão logística no contexto dos eventos raros. Neste sentido foram descritos e comparados um conjunto de modelos de regressão logística.

Os modelos de regressão logística abordados foram: a regressão logística usual, a regressão logística condicional em estudos de caso-controlo, a regressão logística de Firth e o modelo de regressão logística proposto por King e Zeng.

Tendo em conta as especificidades do modelo de regressão logística condicional em estudos de caso-controlo foi possível abordar o tema dos estudos de caso-controlo. Aqui é de destacar o papel das variáveis de confundimento no emparelhamentos de cada caso com um ou mais controlos. Ao nível da aplicação feita para este modelo de regressão logística, a criação dos conjuntos correspondentes terá sido, muito possivelmente, o passo mais importante e difícil. É importante salientar que houve a necessidade de criar um *script* de comandos propositadamente para efetuar os emparelhamentos.

Para o conjunto de dados utilizados nesta aplicação, dados estes que correspondem à informação de todos os diagnósticos de Tuberculose em Portugal entre os anos de 2008 e de 2015, foi possível concluir que o melhor número de controlos a emparelhar com cada caso é cinco ($M = 5$), isto é, foi possível concluir que o conjunto de dados em aplicação devia ser estudado utilizando a correspondência 1:5.

No que é relativo à comparação dos modelos, já sabíamos, logo à partida, que o modelo de regressão logística usual não seria o melhor modelo de regressão para modelar a relação entre uma variável aleatória binária representativa de um evento raro e uma ou mais variáveis explicativas. Tal pressuposto deve-se ao facto de este modelo tendencialmente subestimar a probabilidade da ocorrência do evento raro, sobrestimando a sua não ocorrência, constituindo este pressuposto o principal fundamento para a presente dissertação. Este pressuposto fica confirmado quando observamos a representação gráfica da **Curva ROC** e o valor obtido para a medida de **AUC**, uma vez que a **Curva ROC** resultante da aplicação do modelo de regressão logística usual ao conjunto de dados não é mais do que um segmento de reta, isto é, apresenta uma curvatura nula.

O mesmo resultado se conclui para o modelo de regressão logística de Firth e para o modelo

de regressão logística proposto pelos autores King e Zeng. Com alguma surpresa estes modelos não foram capazes de aprender suficientemente os dados, o que se confirmou, não só olhando para os valores obtido para as várias medidas utilizadas, mas também ao nível das matrizes de confusão resultantes da aplicação dos modelos de regressão na previsão das observações dos conjuntos de teste. No caso da matriz de confusão correspondente à aplicação do modelo de regressão logística proposta por King e Zeng, podemos mesmo ver que este não conseguiu prever um único caso.

Dos quatro modelos aplicados o que apresentou, genericamente, melhores resultados foi o modelo de regressão logística condicional em estudos de caso-controlo . Este modelo foi aplicado ao nível do software **R** utilizando-se a função '*clogit( )*' pertencente à biblioteca **survival**.

Ao nível da representação da **Curva ROC** este foi o único modelo para o qual a curva não é um simples segmento de reta. Ao mesmo tempo, é quando olhamos para a matriz de confusão obtida para este modelo e a comparamos com as matrizes obtidas para os restantes modelos que nos apercebemos o quanto é que o resultado deste modelo é melhor que os demais, simplesmente por ter conseguido prever corretamente mais de 37 % dos casos, ainda que para tal tenha errado na previsão de vários controlos.

Desta feita podemos concluir que o modelo de regressão logística condicional em estudos de caso-controlo foi o que melhor se adaptou aos nossos dados, sendo pois, o que melhor nos permite prever, de entre os modelos estudados, o evento raro correspondente ao conjunto de dados utilizado.

Numa leitura ao melhor modelo obtido de entre todos os estudados, podemos dizer que este se socorre de três variáveis explicativas, que concluímos que sejam as mais significativas na previsão do desfecho do tratamento à Tuberculose em Portugal, tendo em conta os emparelhamentos obtidos. Tais variáveis são: **País de origem**, **HR** e **Comorbilidades**.

É de se notar que o objetivo do modelo de regressão logística condicional em estudos de caso-controlo é diferente do objetivo dos restantes modelos. Tal diferença é essencialmente devida ao facto de o ajustamento do modelo ser feito tendo em conta os emparelhamentos dos casos com os controlos utilizando as variáveis de confundimento, variáveis estas que não constam do modelo final.

Tendo em conta o valor da exponencial dos coeficientes obtidos para cada uma das variáveis, podemos afirmar que:

- 0.651 representa o odds ratio para a morte durante o tratamento à TB (versus sobreviver ao tratamento) para quem nasceu em Portugal, por oposição a quem não nasceu em Portugal. Sendo este valor bastante significativo, podemos dizer que ter nascido em Portugal constitui um fator de proteção para o desfecho do tratamento à TB;
- por seu turno, o valor de 2.511 que representa o odds ratio para a morte durante o tratamento à TB (versus sobreviver ao tratamento) para quem apresenta multirresistência, em contraste com quem não apresenta mutirresistência, permite-nos concluir que a presença de multirresistência é um fator de risco para a morte durante o tratamento à TB, dada a significância deste valor; e

- relativamente ao valor da exponencial do coeficiente obtido para a variável **Comorbilidades**, podemos dizer, tendo em conta a sua significância que a presença de outras doenças, tais como a diabetes, a silicose, a sarcoidose, a neoplasia do pulmão, entre outras, constituem um fator de risco que favorece a morte durante o tratamento à Tuberculose em Portugal.

Em modo de conclusão, e ainda que os resultados não tenham sido extraordinários, penso que os objetivos propostos para a presente dissertação foram cumpridos, uma vez que os vários objetivos elencados inicialmente foram abordados.

8.1 Trabalhos futuros

Em termos de trabalhos futuros penso que existem várias linhas de interesse que podem ser seguidas, ainda que meras gotas no vasto oceano que é a regressão logística. Algumas destas linhas de interesse foram pouco esmiuçadas na presente dissertação. De seguida, enumero três dessas potenciais linhas de interesse:

i. uma abordagem mais profunda dos estudos da regressão logística de Firth e do modelo de regressão logística proposto por King e Zeng;

ii. utilizar dados simulados para a comparação dos vários modelos de regressão; e

iii. avaliar as medidas que melhor nos permitem comparar modelos de regressão para dados com eventos raros.

Bibliografia

[Agresti, 2002] Agresti, Alan (2002) **Categorical data analysis** (2nd ed) *John Wiley & Sons*: New Jersey.

[Baltar & Okano, 2017] **Análise de Concordância - Kappa**. Acedido a 19 de agosto de 2017, em: `http://www.lee.dante.br/pesquisa/kappa/`.

[Bonita & et al., 2006] Bonita, R., Beaglehole, R. & Kjellöm, T. (2006) **Basic Epidemiology**. (2nd ed) *World Health Organization.*

[Breslow, 1996] Breslow, N. E. (1996) **Statistics in Epidemiology: The Case-Control Study**. *Journal of the American Statistical Association.* Volume 91, Número 433, 14-28.

[Breslow & Day, 1980] Breslow, N. E. & Day, N. E. (1980) **Statistical Methods in Cancer Research. Volume I - The Analysis of CaseControl Studies**. *International Agency for Research on Cancer (IARC Scientific Publications No. 32)*: Lyon.

[CDC, 2017] **Centers for Disease Control and Prevention**. Acedido a 20 de maio de 2017, em: `https://www.cdc.gov/tb/`.

[Clayton & Hills, 2013] Clayton, David & Hills, Michael (2013) **Statistical models in Epidemiology**. *Oxford University Press*: Oxford.

[Collett, 2003] Collett, David (2003) **Modelling Binary Data** (2nd ed) *Chapman & Hall/-CRC*: Boca Raton (Flórida).

[Cox, 1958] Cox, D. R. (1958) **The Regression Analysis of Binary Sequences**. *Journal of the Royal Statistical Society. Serie B (Methodological)* Volume 20, Número 2, 215-241.

[Cox & Hinkley, 1974] Cox, D. R. & Hinkley (1974) **Theoretical statistics**. *Chapman & Hall*: London.

[Crisóstomo & Soares, 2017] Crisóstomo, Pedro & Soares, Rosa (2017, 27 de março). **Incumprimento baixa mas ainda há 134 mil a falhar a prestação da casa**. *Público (online)*. Acedido a 3 de abril de 2017, em: `https://www.publico.pt/2017/03/27/economia/noticia/incumprimento-baixa-mas-ainda-ha-134-mil-a-falhar-a-prestacao-da-casa-1766476`.

[DGS, 2015] **Portugal em números 2015 – Infeção VIH, SIDA e Tuberculose**. Acedido a 22 de maio de 2017, em: `http://www.dgs.pt/?cr=29118`.

[Dobson, 2002] Dobson, Annette J. (2002) **A Introdution to Generalized Linear Models** (2nd ed) *Chapman & Hall/CRC*: Boca Raton (Flórida).

[Firth, 1993] Firth, David (1993) **Bias reduction of maximum likelihood estimates**. *Biometrika* Volume 80, Número 1: 27 – 38.

[Freund & Minton, 1979] Freund, Rudolf J. & Minton, Paul D. (1979) **Regression methods: a tool for data analysis**. *Marcel Dekker*: New York.

[FPP, 2017] **Fundação Portuguesa do Pulmão**. Acedido a 20 de maio de 2017, em: `http://www.fundacaoportuguesadopulmao.org/tuberculose.html`.

[Harrell, 2001] Harrell, Frank E. (2001) **Regression Modeling Strategies - With Applications to Linear Models, Logistic Regression, and Survival Analysis**. *Springer-Verlag*: New York.

[Heinze & Schemper, 2002] Heinze, George & Schemper, Michael (2002) **A solution to the problem of separation in logistic regression**. *Statistics in Medicine*. Capítulo 21: 2409 – 2419.

[Hosmer & Lemeshow, 2000] Hosmer, David & Lemeshow, Stanley (2000) **Applied Logistic Regression**.*Wiley Series in Probability and Statistics*. Capítulo 1: 01 - 10.

[King & Zeng, 2001a] King, Gary & Zeng, Langche (2001) **Logistic Regression in Rare Events Data**. *Political Analysis* Volume 9, Número 2: 137 - 163.

[King & Zeng, 2001b] King, Gary & Zeng, Langche (2001) **Explaining Rare Events in International Relations**. *International Organization* Volume 55, Número 3: 693 – 715.

[Lança, 2003] Lança, Mário Ataíde (2003) **Grande Enciclopédia Médica - Saúde da Família**. *Edição e Conteúdos*: Matosinhos. Volume 14 (66 - 71).

[Lehmann, 1988] Lehmann, Erich L. (1988) **Statistics: An Overview**. *Enc. Of Stat. Science - Wiley*. Volume 8.

[Liddell & et al., 1977] Liddell, F. D. K., McDonald, J. C. & Thomas, D. C. (1977) **Methods of cohort analysis: appraisal by application to asbestos mining**. *J. R. stat. Soc. Ser. A*. Volume 140 (469-491).

[Lobo & et al., 2008] Lobo, Jorge M., Jiménez-Valverde, Alberto & Real, Raimundo (2008) **AUC: a misleading measure of the performance of predictive distribution models**. *Global Ecology and Biogeography*. Volume 17 (145-151).

[McCullagh, 1987] McCullagh, P. (1987) **Tensor Methods in Statistics** *Chapman & Hall*: London.

[McCullagh & Nelder, 1989] McCullagh, P. & Nelder, J. A. (1989) **Generalized Linear Models** (2nd ed) *Chapman & Hall*: London.

[Murteira & et al., 2007] Murteira, B. & Ribeiro, C. S. & Silva, J. A. & Pimenta, C. (2007) **Introdução à estatística** (2nd ed) *McGraw-Hill*: Madrid.

[Nau, 2014] Nau, Robert (2014) **Notes on linear regression analysis**. Acedido a 23 de março de 2017, em: `http://people.duke.edu/~rnau/Notes_on_linear_regression_analysis--Robert_Nau.pdf`.

[Oliveira & Massamo, 2012] Oliveira, Ana & Massamo, João (2012) **Síndrome de Gilles de La Tourette: Clínica, diagnóstico e abordagem terapêutica**. *Arq Med.* Volume 26 (211-217).

[R Project, 2017] **The R Project for Statistical Computing**. Acedido a 19 de maio de 2017, em: `https://www.r-project.org/`.

[Shapiro, 1954] Shapiro, Helene (1954) **Linear Algebra and Matrices - Topics for a Second Course**. *American Mathematical Society*: Providence, Rhode Island.

[Snipes & Taylor, 2014] Snipes, Michael & Taylor, Christopher (2014) **Model selection and Akaike Information Criteria: An example from wine ratings and prices**. *Wine Economics and Policy - ELSEVIER* Volume 3 (3–9).

[Song & Chung, 2010] Song, Jae W. & Chung, Kevin C. (2010) **Observational Studies: Cohort and Case-Control Studies**. *Plast Reconstr Surg* Volume 126 (2234–2242).

[Stare & Maucort-Boulch, 2016] Stare, Janez & Maucort-Boulch, Delphine (2016) **Odds Ratio, Hazard Ratio and Relative Risk**. *Metodoloski zvezki.* Volume 13 (59-67).

[Turkman & Silva, 2000] Turkman, Maria Antónia & Silva, Giovani (2000) **Modelos Lineares Generalizados - da teoria à prática**. *UL & UTL*: Lisboa.

[Wilson & Lorenz, 2015] Wilson, Jeffrey R. & Lorenz, Kent A. (2015) **Modeling Binary Correlated Responses using SAS, SPSS and R** *Springer*: New York.

[WHO, 2016] **Global tuberculosis report 2016**. Acedido a 20 de maio de 2017, em: `http://www.who.int/en/`.

Anexos

Anexo A - Descrição das variáveis do conjunto de dados

Na tabela que se segue apresenta-se uma breve análise descritiva das variáveis do conjunto de dados inicial, já sem as observações cujo **Outcome** é **ET-TF**.

Variável	Total	Outcome	
		Bom outcome	Mau outcome
Sexo	n (%)	n (%)	n (%)
Masculino	12781 (65.480)	11240 (64.450)	1541 (74.122)
Feminino	6738 (34.520)	6200 (35.550)	538 (25.878)
Idade	Md [1] (min - max)	Md (min - max)	Md (min - max)
	45 (0 - 100)	45 (0 - 100)	55 (0 - 98)
País de origem	n (%)	n (%)	n (%)
Portugal	16509 (84.579)	14758 (84.622)	1751 (84.223)
Angola	679 (3.479)	598 (3.429)	81 (3.896)
Guiné-Bissau	515 (2.638)	455 (2.609)	60 (2.886)
Cabo Verde	489 (2.505)	434 (2.489)	55 (2.646)
Brasil	248 (1.271)	236 (1.353)	12 (0.577)
Moçambique	193 (0.989)	168 (0.963)	25 (1.203)
Roménia	139 (0.712)	124 (0.711)	15 (0.722)
S. Tomé e Príncipe	106 (0.543)	92 (0.528)	14 (0.673)
Outros [2]	641 (3.284)	575 (3.297)	66 (3.175)
Histórico de Reclusão	n (%)	n (%)	n (%)
0 (Não)	19196 (98.345)	17147 (98.320)	2049 (98.557)
1 (Sim)	323 (1.655)	293 (1.680)	30 (1.443)
Rastreio	n (%)	n (%)	n (%)
Rastreio passivo	17639 (90.368)	15728 (90.183)	1911 (91.919)
Outro rastreio	1148 (5.881)	1080 (6.193)	68 (3.271)
Desconhecido	732 (3.750)	632 (3.624)	100 (4.810)

[1] Não existindo uma notação amplamente definida para a **mediana**, usaremos **Md** para este efeito.

[2] **Outros** reúne as contagens dispersas que valem menos de 0.500% da totalidade das observações.

Tempo para a primeira consulta	Md (min - max)	Md (min - max)	Md (min - max)
	34 (0 - 7536)	34 (0 - 4527)	31 (0 - 7536)
Insuficiência renal crónica em diálise	n (%)	n (%)	n (%)
Não	19270 (98.724)	17259 (98.962)	2011 (96.729)
Sim	249 (1.276)	181 (1.038)	68 (3.271)
Linfomas ou Doenças Mieloproliferativas	n (%)	n (%)	n (%)
Não	19407 (99.426)	17363 (99.558)	2044 (98.316)
Sim	112 (0.574)	77 (0.442)	35 (1.684)
Infeção por VIH	n (%)	n (%)	n (%)
Não	17374 (89.011)	15751 (90.315)	1623 (78.066)
Sim	2145 (10.989)	1689 (9.685)	456 (21.934)
Inflamatória Articular	n (%)	n (%)	n (%)
Não	19339 (99.078)	17274 (99.048)	2065 (99.327)
Sim	180 (0.922)	166 (0.952)	14 (0.673)
Outra doença do Interstício	n (%)	n (%)	n (%)
Não	19468 (99.739)	17393 (99.731)	2075 (99.808)
Sim	51 (0.261)	47 (0.269)	4 (0.192)
Diabetes	n (%)	n (%)	n (%)
Não	18328 (93.898)	16437 (94.249)	1891 (90.957)
Sim	1191 (6.102)	1003 (5.751)	188 (9.043)
Silicose	n (%)	n (%)	n (%)
Não	19276 (98.755)	17225 (98.767)	2051 (98.653)
Sim	243 (1.245)	215 (1.233)	28 (1.347)
Doença hepática	n (%)	n (%)	n (%)
Não	18717 (95.891)	16803 (96.347)	1914 (92.063)
Sim	802 (4.109)	637 (3.653)	165 (7.937)
Neoplasia do pulmão	n (%)	n (%)	n (%)
Não	19359 (99.180)	17361 (99.547)	1998 (96.104)
Sim	160 (0.820)	79 (0.453)	81 (3.896)
Sarcoidose	n (%)	n (%)	n (%)
Não	19490 (99.851)	17414 (99.851)	2076 (99.856)
Sim	29 (0.149)	26 (0.149)	3 (0.144)
Neoplasia de outros órgãos	n (%)	n (%)	n (%)
Não	18907 (96.865)	16988 (97.408)	1919 (92.304)

Sim	612 (3.135)	452 (2.592)	160 (7.696)
DPOC	n (%)	n (%)	n (%)
Não	18965 (97.162)	17002 (97.489)	1963 (94.420)
Sim	554 (2.838)	438 (2.511)	116 (5.580)
Depend. de Álcool	n (%)	n (%)	n (%)
Não	16082 (82.392)	14624 (83.853)	1458 (70.130)
Sim	2386 (12.224)	2008 (11.514)	378 (18.182)
Desconhecido	1051 (5.384)	808 (4.633)	243 (11.688)
Dependência de drogas injetáveis	n (%)	n (%)	n (%)
Não	17313 (88.698)	15727 (90.170)	1586 (76.287)
Sim	1202 (6.158)	939 (5.400)	263 (12.650)
Desconhecido	1004 (5.144)	774 (4.430)	230 (11.063)
Principal localização da doença	n (%)	n (%)	n (%)
Pulmonar	14082 (72.145)	12562 (72.030)	1520 (73.112)
Pleural	1640 (8.402)	1479 (8.481)	161 (7.744)
Linfática extratoracica	1342 (6.875)	1240 (7.110)	102 (4.906)
Linfática intratoracica	297 (1.522)	271 (1.554)	26 (1.251)
Genito/Urinária	364 (1.865)	327 (1.875)	37 (1.780)
Vertebral	299 (1.532)	265 (1.519)	34 (1.635)
Disseminada	223 (1.142)	176 (1.009)	47 (2.261)
Peritoneal/Digestiva	199 (1.020)	173 (0.992)	26 (1.251)
Meningite	144 (0.738)	112 (0.642)	32 (1.539)
Osteoartic. N. Verteb.	118 (0.605)	108 (0.619)	10 (0.481)
SNC	52 (0.266)	42 (0.241)	10 (0.481)
Outra	676 (3.463)	611 (3.503)	65 (3.127)
Desconhecida	83 (0.425)	74 (0.424)	9 (0.432)
Raio-X Torax	n (%)	n (%)	n (%)
Cavitada	6988 (35.800)	6385 (36.611)	603 (29.004)
Não Cavitada	8065 (41.319)	7115 (40.797)	950 (45.695)
Normal	2934 (15.032)	2696 (15.459)	238 (11.448)
Desconhecida	1532 (7.849)	1244 (7.133)	288 (13.853)
N. tratamentos anteriores	n (%)	n (%)	n (%)
0	17822 (91.306)	15991 (91.692)	1831 (88.071)
> 0	1697 (8.694)	1449 (8.308)	248 (11.929)
ANO	n (%)	n (%)	n (%)
2008	2881 (14.760)	2596 (14.885)	285 (13.709)
2009	2770 (14.191)	2493 (14.295)	277 (13.324)
2010	2592 (13.279)	2305 (13.217)	287 (13.805)

2011	2526 (12.941)	2284 (13.096)	242 (11.640)
2012	2494 (12.777)	2252 (12.913)	242 (11.640)
2013	2242 (11.486)	1990 (11.411)	252 (12.121)
2014	2090 (10.708)	1839 (10.545)	251 (12.073)
2015	1924 (9.857)	1681 (9.639)	243 (11.688)
HR	n (%)	n (%)	n (%)
0	19345 (99.109)	17307 (99.237)	2038 (98.028)
1	174 (0.891)	133 (0.763)	41 (1.972)

Tabela 8.1: Breve descrição das variáveis que fazem parte do conjunto de dados.

Anexo B - 'Sumário' do modelo *glm1*

Neste anexo é relatado o resultado obtido no software **R**, utilizando o comando '*summary(glm1)*'.

```
> summary(glm1);

Call:
glm(formula = I(Outcome) ~ I(Sexo) + I(Idade) + I(Pais.de.Origem) +
    I(Dep..de.alcool) + I(ANO) + I(Hx.reclusao) + I(Infeccao.por.VIH) +
    I(Dep..drogas.ev) + I(Rastreio) + Tempo.para.a.primeira.consulta +
    I(TB.Doenca.Localizacao.Principal) + I(RX.Torax) + I(N..tratamentos.anteriores) +
    I(HR) + I(Comorbilidades), family = binomial(link = "logit"),
    data = DTreino)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4376  -0.4885  -0.3193  -0.2572   2.9527

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                             -3.4211793  0.3757668  -9.105  < 2e-16 ***
I(Sexo)1                                 0.2375187  0.0837691   2.835 0.004577 **
I(Idade)>60                              1.7184818  0.2814058   6.107 1.02e-09 ***
I(Idade)20 - 40                          0.4638908  0.2840516   1.633 0.102443
I(Idade)40 - 60                          0.6261993  0.2827317   2.215 0.026773 *
I(Pais.de.Origem)Portugal               -0.3370132  0.1003545  -3.358 0.000784 ***
I(Dep..de.alcool)1                       0.5804590  0.0977605   5.938 2.89e-09 ***
I(Dep..de.alcool)Desconhecido            0.7649559  0.1723497   4.438 9.06e-06 ***
I(ANO)2014+                              0.0407884  0.0898632   0.454 0.649904
I(Hx.reclusao)1                         -0.6818357  0.3551413  -1.920 0.054871 .
I(Infeccao.por.VIH)1                     1.0336872  0.1071702   9.645  < 2e-16 ***
I(Dep..drogas.ev)1                       0.7719351  0.1352724   5.707 1.15e-08 ***
I(Dep..drogas.ev)Desconhecido            0.4417109  0.1864223   2.369 0.017816 *
I(Rastreio)Outro_rastreio               -0.8012782  0.3382823  -2.369 0.017852 *
I(Rastreio)Rastreio_passivo             -0.2164652  0.1734063  -1.248 0.211917
Tempo.para.a.primeira.consulta          -0.0008031  0.0004284  -1.875 0.060844 .
I(TB.Doenca.Localizacao.Principal)Pleural -0.0505464  0.1701510  -0.297 0.766415
I(TB.Doenca.Localizacao.Principal)Pulmonar  0.1708089  0.1279477   1.335 0.181880
I(RX.Torax)Desconhecida                  0.7025311  0.1430950   4.910 9.13e-07 ***
I(RX.Torax)Nao Cavitada                  0.1646691  0.0848255   1.941 0.052226 .
I(RX.Torax)Normal                       -0.1172861  0.1634233  -0.718 0.472953
I(N..tratamentos.anteriores)0           -0.1231386  0.1111196  -1.108 0.267792
I(HR)1                                   0.8247735  0.2778591   2.968 0.002994 **
I(Comorbilidades)1                       0.6851942  0.0871054   7.866 3.65e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6511.4  on 10195  degrees of freedom
Residual deviance: 5798.0  on 10172  degrees of freedom
AIC: 5846

Number of Fisher Scoring iterations: 6
```

Anexo C - Script com o conjunto de instruções a efetuar os emparelhamentos caso-controlo

Neste anexo é apresentado o *script* criado para efetuar os emparelhamentos entre os casos e os controlos.

```
### o------------------o Conjuntos correspondentes: o------------------o
MControls = 1;

XYZ <- cbind(Base1$Sexo, Base1$Idade, Base1$Infeccao.por.VIH,
             Base1$Dep..de.alcool, Base1$Dep..drogas.ev)
IdMissing <- which(!complete.cases(XYZ)); Base1X <- Base1[-IdMissing, ];
ID <- c(1:length(Base1X[,1])); Data <- cbind(ID, Base1X);
rm(XYZ, IdMissing, Base1X, ID)

### Emparelhamento
Data1<-Data[Data$Outcome==1, ]; Data0<-Data[Data$Outcome==0, ];

set.seed(31); i = 0; NoMatche <- c(); contSetID = 0;
DataC <- data.frame(SetId = contSetID, IdCC = 0, Data1[1,], row.names = i)
for(j in 1:length(Data1[,1])){
  i = i + 1;
  var1 <- Data1$Sexo[j]; var2 <- Data1$Infeccao.por.VIH[j];
  var3 <- Data1$Dep..de.alcool[j]; var4 <- Data1$Dep..drogas.ev[j];
  var5 <- Data1$Idade[j];
  DataX <- Data0[Data0$Sexo==var1 & Data0$Infeccao.por.VIH==var2 &
                 Data0$Dep..de.alcool==var3 & Data0$Dep..drogas.ev==var4 &
                 Data0$Idade==var5, ]

  if(length(DataX[,1]) >= MControls){
    contSetID = contSetID + 1;
    X <- data.frame(SetId = contSetID, IdCC = 0, Data1[j, ], row.names = i);
    DataC <- rbind(DataC, X); contControl = 0;
    while(contControl < MControls){
      p <- round(runif(1, 1, length(DataX[,1])))
      contControl = contControl + 1; i = i + 1;
      Y <- data.frame(SetId = contSetID, IdCC = contControl,
                      DataX[p, ], row.names = i);
      DataC <- rbind(DataC, Y)
      Data0 <- Data0[!(Data0$ID == DataX[p,1]), ];
      DataX <- DataX[-p, ];     } }
  else{NoMatche <- c(NoMatche, Data1[j,1])}
  DataCC <- DataC[-1, ]; DataCC <- DataCC[,-3]; }
```

Anexo D - Resultados intermédios: regressão logística condicional em estudos de caso-controlo

Neste anexo são apresentados os vários resultados intermédios da aplicação da regressão logística condicional em estudos de caso-controlo , tendo em conta os vários valores de **M**.

- $M = 1$:

Modelo	**Valor - p**	**AIC**	**Acurácia**	**Kappa**	**AUC**
clogit11	< 0.001	984.986	0.520	0.061	0.532
clogit12	< 0.001	981.256	0.520	0.061	0.531
clogit13	< 0.001	979.345	0.518	0.058	0.530
clogit14	< 0.001	**977.512**	**0.521**	**0.063**	**0.532**
clogit15	< 0.001	977.822	0.530	0.051	0.526
clogit16	< 0.001	976.967	0.523	0.026	0.513
clogit17	< 0.001	977.516	0.532	0.041	0.520

Tabela 8.2: Valores de algumas medidas dos modelos avaliados com a regressão logística condicional em estudos de caso-controlo , tendo em conta o emparelhamento 1:1.

- $M = 2$:

Modelo	**Valor - p**	**AIC**	**Acurácia**	**Kappa**	**AUC**
clogit21	< 0.001	1616.693	0.476	0.043	0.528
clogit22	< 0.001	1614.915	0.476	0.043	0.528
clogit23	< 0.001	1613.163	0.477	0.045	0.529
clogit24	< 0.001	1611.221	0.477	0.045	0.529
clogit25	< 0.001	1608.111	0.477	0.044	0.528
clogit26	< 0.001	**1623.814**	**0.601**	**0.070**	**0.535**
clogit27	< 0.001	1624.395	0.600	0.069	0.535
clogit28	< 0.001	1624.395	0.600	0.069	0.535

Tabela 8.3: Valores de algumas medidas dos modelos avaliados com a regressão logística condicional em estudos de caso-controlo , tendo em conta o emparelhamento 1:2.

- $M = 3$:

Modelo	**Valor - p**	**AIC**	**Acurácia**	**Kappa**	**AUC**
clogit31	< 0.001	2076.628	0.444	0.051	0.541
clogit32	< 0.001	2074.662	0.444	0.050	0.541
clogit33	< 0.001	2071.224	0.442	0.048	0.539
clogit34	< 0.001	**2069.650**	**0.443**	**0.049**	**0.540**
clogit35	< 0.001	2069.247	0.465	0.049	0.538
clogit36	< 0.001	2076.223	0.272	0.008	0.508
clogit37	< 0.001	2080.100	0.253	0.005	0.505

Tabela 8.4: Valores de algumas medidas dos modelos avaliados com a regressão logística condicional em estudos de caso-controlo , tendo em conta o emparelhamento 1:3.

- $M = 4$:

Modelo	**Valor - p**	**AIC**	**Acurácia**	**Kappa**	**AUC**
clogit41	< 0.001	2407.073	0.429	0.041	0.539
clogit42	< 0.001	2405.075	0.430	0.042	0.540
clogit43	< 0.001	2402.649	0.448	0.044	0.540
clogit44	< 0.001	2400.770	0.448	0.044	0.540
clogit45	< 0.001	2399.568	0.448	0.044	0.540
clogit46	< 0.001	**2408.398**	**0.627**	**0.063**	**0.540**
clogit47	< 0.001	2408.595	0.626	0.062	0.539
clogit48	< 0.001	2412.527	0.636	0.062	0.538

Tabela 8.5: Valores de algumas medidas dos modelos avaliados com a regressão logística condicional em estudos de caso-controlo , tendo em conta o emparelhamento 1:4.

- $M = 5$:

Modelo	**Valor - p**	**AIC**	**Acurácia**	**Kappa**	**AUC**
clogit51	< 0.001	2509.457	0.436	0.036	0.539
clogit52	< 0.001	2506.006	0.436	0.037	0.540
clogit53	< 0.001	2504.194	0.438	0.035	0.538
clogit54	< 0.001	2502.732	0.439	0.035	0.537
clogit55	< 0.001	2501.710	0.448	0.039	0.541
clogit56	< 0.001	2509.131	0.194	0.004	0.505
clogit57	< 0.001	2508.845	0.648	0.066	0.546
clogit58	< 0.001	**2510.472**	**0.661**	**0.068**	**0.546**

Tabela 8.6: Valores de algumas medidas dos modelos avaliados com a regressão logística condicional em estudos de caso-controlo , tendo em conta o emparelhamento 1:5.